FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Data Analytics Mechanisms for Sports' Related Data Retrieval and Multimedia Augmentation

Tiago André Pérola Filipe

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Maria Teresa Magalhães da Silva Pinto de Andrade

Second Supervisor: Maria Carvalho Goreti Marreiros

July 03, 2019

Data Analytics Mechanisms for Sports' Related Data Retrieval and Multimedia Augmentation

Tiago André Pérola Filipe

Mestrado Integrado em Engenharia Informática e Computação

July 03, 2019

Abstract

For many years, sports have been and are still a part of our lives, whether as players or as spectators. In this context, the diffusion and massification of platforms for multimedia content visualization have made the consumption of these same contents widespread.

In the meantime, there is more and more information available about clubs, players, matches, and competition venues, which is used in addition to the game itself. Thus, it is relevant to apply mechanisms to collect these data and extract valuable information from it, such as the odds of a team winning a game, or knowing what team is best suited to play against the opponent. The next step is to present this information, in which the target audience is composed of sports organization's staff and the spectators. For the staff, the information is in the form of reports, and for the spectators is presented as an overlay above the game transmission, without there being a perceptible intrusion on the main content.

Throughout this dissertation, it was implemented a full business intelligence pipeline in order to extract more value from the collected data. The data collected is from soccer and basketball, more precisely from the National Basketball Association (NBA) and the English Premier League (EPL). After the integration of data into a multidimensional database, several layers of analysis were applied to the data: Business Intelligence (BI) Reporting, Machine Learning (ML), and Simulation. The results of these methodologies were used to feed the multimedia augmentation layer.

BI allowed an exploratory data analysis, which was the source of content for the overlays. ML was applied to predict game outcomes, in which the algorithms used were: decision tree, support vector machine, random forest, extremely randomized trees, and extreme gradient boosting. The best accuracy was achieved with the SVM for basketball and with the extreme gradient boosting for soccer, with values of 67.96% and 54.44% respectively. In order to forecast outcomes a basketball simulation system was developed. This system allows for play-by-play simulation of a game and achieved a performance of 64.80%. The system also supports the positioning of various informational overlays, with game statistics, on a video stream, either live or on demand.

The ML models exceeded the performance of many works found in the literature, which solidifies the potential impact of the dissertation work herein developed. Furthermore, simulation creates value to the state of the art, since few articles describe in detail the development of a sports simulator.

Resumo

Durante muitos anos, o desporto fez e continua a fazer parte das nossas vidas, seja como jogadores ou como espectadores. Neste contexto, a difusão e massificação de plataformas para visualização de conteúdo multimédia alargou o consumo destes mesmos conteúdos.

Enquanto isso, há mais e mais informações disponíveis sobre clubes, jogadores, jogos e locais de competição, que são usados como adição ao próprio jogo. Deste modo, é relevante aplicar mecanismos para armazenar esses dados e extrair informação útil deles, como a probabilidade de uma equipa vencer um jogo, ou se saber qual a equipa mais adequada para jogar contra o adversário. O próximo passo é apresentar essa informação, que tem como público-alvo o staff das organizações desportivas e os espectadores. Para o staff é na forma de relatórios, e para os espectadores na forma de *overlay* por cima da transmissão do jogo, sem que haja uma intrusão percetível no conteúdo principal.

Nesta dissertação foi implementada uma *pipeline* de *business intelligence* com a intenção de extrair maior valor dos dados armazenados. Os dados armazenados são de futebol e basquetebol, mais precisamente da National Basketball Association (NBA) e da English Premier League (EPL). Após a integração dos dados numa base de dados multidimensional, várias camadas de análise foram aplicadas aos dados: *Business Intelligence* (BI) *Reporting*, *Machine Learning* (ML) e simulação. Os resultados destas metodologias foram utilizados para construir uma base de informação de suporte à camada multimédia.

BI permitiu uma análise exploratória dos dados, que foi a fonte do conteúdo dos overlays. ML foi aplicado para prever os resultados dos jogos, em que os algoritmos foram: *decision tree*, *support vector machine*, *random forest*, *extremely randomized trees* e *extreme gradient boosting*. A melhor *accuracy* foi obtida com o modelo SVM para o basquetebol e com o *extreme gradient boosting* para o futebol, com valores de 67.96% e 54.44% respetivamente. Com o objetivo de prever resultados foi desenvolvido um simulador para basquetebol. Este simulador permite simular um jogo jogada a jogada e atingiu uma performance de 64.80%. O sistema desenvolvido também permite posicionar vários *overlays* informativos, com estatísticas dos jogos, por cima do vídeo da transmissão, quer seja ao vivo ou sob demanda.

Os modelos de ML excederam a performance de vários trabalhos encontrados na literatura, o que solidifica o potencial impacto do trabalho desta dissertação desenvolvida. Para além disso, a simulação cria valor ao estado de arte, porque poucos artigos descrevem detalhadamente o desenvolvimento de um simulador desportivo.

Acknowledgements

This dissertation would not have been possible without the precious help of several people.

First, I would like to thank my supervisors, Prof. Maria Teresa Andrade, and Prof. Goreti Marreiros, for accepting this challenge, guiding me along this path always with confidence.

I am thankful to MOG Technologies, for providing good conditions to develop this work, especially to Nuno Cravino, who supervised me inside the company and at the same time helped me during the development of the dissertation, to Alexandre Ulisses and Ivone Amorim, for the challenge proposed and the confidence given, and to the co-workers that helped and shared experiences with me during this time.

Finally, a very special thanks to my family who always believed in me and provided me with the best conditions to fight and achieve my life goals, and to my girlfriend, Patricia, who has always been with me in the good and the bad times.

Tiago André Pérola Filipe

"Don't ever let someone tell you that you can't do something."

Will Smith (The Pursuit of Happyness, movie)

Contents

List of Figures

List of Tables

Abbreviations

API	Application programming interface
BI	Business Intelligence
BI&A	Business Intelligence & Analytics
C	Center
CSV	Comma-Separated Values
EPL	English Premier League
HLS	HTTP Live Streaming
HMM	Hidden Markov Model
HTML	HyperText Markup Language
LSTM	Long short-term memory
MDP	Markov decision process
ML	Machine Learning
MLB	Major League Baseball
NBA	National Basketball Association
NFL	National Football League
PF	Power Forward
PG	Point Guard
POMDP	Partially observable Markov decision process
RBF	Radial Basis Kernel
REST	Representational State Transfer
RF	Random Forest
SF	Small Forward
SG	Shooting Guard
SVM	Support Vector Machine
URL	Uniform Resource Locator
XGBoost	Extreme Gradient Boosting

Chapter 1

Introduction

Sports analytics has been a trend that has grown a lot and remains extremely popular. The MIT Sloan Sports Analytics Conference, a conference founded in 2006, is one of the most prestigious conferences for the creation and diffusion of knowledge about sports analytics, and over these years remains very popular. This year's edition (13^{th}) had 3,500 participants from 33 countries, 44 U. S. states, 130 professional teams, and around 200 universities. [1].

Concerning fan consumption, an important fact is that more than 60% of adult US men and 40% US women, who watch TV will regularly watch sports on TV in 2019. The predictions report written by Deloitte [2] also reveals that "TV sports will represent about two-thirds of all TV watching among 18-24-year-old-men, and more than three-quarters of all TV watching for men age 25-34 who watch TV sports". Additional detail is that more than 40% of 18-34-year-old US men will tend to bet on sports weekly or more often when watching TV sports. Therefore, spectators need to be provided with some meaningful insights.

Enhancement of the viewer experience, plus the insights from predictive models is the perfect match to bring more spectators to sports and keep those who follow.

1.1 Context

Currently, there is a wealth of data on clubs, players, and sports games. The distribution and massification of platforms for multimedia content visualization have promoted the consumption of these contents. As a result, sports data analysis is becoming extensive and diversified. In 2018, the Sports Analytics Market was valued at USD 0.56 billion and by 2022 is estimated to reach a CAGR of 30% over the forecast period of 2019-2014 [3].

Data analytics in sports has reached an important level where actions are no longer based only on the experience and intuition of a player or coach. Proof of this was the Major League Baseball's achievement from the team Houston Astros, who not only won the championship for the first time but made a walk that allowed them to move from last to first. This 2017 triumph was a process

of years that relied heavily on advanced data analytics. Since 2011, the club general manager Jeff Luhnow and his team had to build an organization and culture around data, from that point analytic insight fuelled both player selection and on-the-field decision making, such as where to position players in game situations [4].

This dissertation was carried out at the company MOG Technologies, located in Maia. MOG main market are broadcasters and content producers worldwide, including sports related corporations. Its products range from post-production tools to streaming platforms for live and on demand content. In 2017 it started a new AI and Data Science department that focus on generating new media-related insights from the operational level and all the way up to content delivery.

This project is included in that department and is one of the company's first projects related to the application of data analysis in sports.

1.2 Motivation

Technological evolution has brought easy access to a large amount of data that needs to be processed as quickly as possible to provide useful information.

Recently, almost every industry has started to rely on big data analytics to support decision-making processes. Sports follows the trend by, per example, mitigate in-game injuries and build predictive models to identifying optimal shooting zones from a National Basketball Association (NBA) team [5].

Professional sports teams, such as the Houston Astros, the Chicago Cubs, and the Toronto Maple Leafs, built analytics department and consequently, this tends to be followed by other teams.

Professional sports teams, such as the Houston Astros, the Chicago Cubs, and the Toronto Maple Leafs, built analytics department, which consequently tends to be followed by other teams.

On the other hand, there is the example of the NBA player Kevin Durant who hired a statistical expert to improve his in-game performance [6].

1.3 Objetives

The primary goal of this dissertation is to develop an information retrieval system that is able to increase the sports fan experience and support broadcasters and clubs with meaningful insights.

This system is intended to collect sports related data, such as information from clubs, players, matches, and competition venues, and apply data analytics mechanisms to extract relevant information.

Additionally, the system will offer a multimedia augmentation layer, in order to enhance the sports fan experience, the purpose is to give additional information, without there being a perceptible direct intrusion on the main content. The information will be presented as an overlay above the video containing statistical facts, like the athletes' performance in a given month.

In order to provide meaningful insights to both clubs and broadcasters, this development will offer the ability to analyze the odds of a team winning a game, and be able to help determine which lineups are better against each opponent.

Finally, it is intended that the system covers basketball and soccer.

1.4 Document Structure

The structure of the document is composed of the current chapter and six more chapters.

Chapter 2, analyzes various mechanisms of data analytics in sports. These mechanisms are divided into Machine Learning, Business Intelligence, and simulation. In each mechanism is shown the related work and conclusions about each subsection.

Chapter 3, describes the usage of multimedia analytics in sports. Whether improving the viewer's experience or by analyzing video that aids athletes and coaches. It presents related work and existing products.

Chapter 4, explains how the data from NBA and EPL was collected and organized into a multidimensional database.

Chapter 5, describes the mechanisms implemented to predict game outcomes, which is divided into ML and simulation.

Chapter 6, presents the proposed solution to improve the sports fan's experience and provide knowledge to broadcasters and clubs, whether for soccer or for basketball.

The document ends with Chapter 7, which gives a summary of the work developed, and future work.

Chapter 2

Data Analytics in Sports

It is common to say that baseball was the starting point of sports analytics, where experts have used advanced statistics to improve player selection and game strategy. First, the book "Percentage Baseball" [7] published in 1964 led to the development of a new concept entitled sabermetrics. Sabermetrics is related to the acronym SABR, which means Society for American Baseball Research and was defined by Bill James. Some years later, emerged the popular movie Moneyball [8], in which sabermetrics and advanced statistics helped Oakland Athletics in decision making.

Since then, the diversity of statistical technologies and techniques lead to an improvement in data collection and decision making in competitive sports.

2.1 Business Intelligence

Decision-making processes have a tremendous impact on the success of an organization. Decision makers, namely top managers, usually used the knowledge acquired with experience to decide. During the last years, Business Intelligence (BI) has appeared as an important topic of Information Systems. Consequently, companies investing in it rely on data with the purpose of getting summarize information in an organized manner that supports the strategic decision [9, Chapter 4].

In Sports, curiosities about two matchups, for instance, how many times both times played against each other, statistics from a player in the current month, and what is the average age of a team are examples of tasks that can be calculated through BI.

The development of data warehousing concepts and techniques, the evolution of data visualization techniques and the production of reports and dashboards are key drivers of BI [9, Chapter 4]. These key drivers and BI existing products are described in the next chapters.

2.1.1 Data Warehouse

In the last years, with the emergence of Big Data, several changes have occurred at the data warehouse level, and everything indicates that in the future more changes will occur. Big Data

enabled more and more organizations to focus on storing the vast amount of data available to gain business advantage. Facebook and Google are good examples of organizations that provide a wealth of data for the entire world. [10, Chapter 1]

A data warehouse can be defined as "a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making." [11, Chapter 1]

Its production involves a technical component and a business component. In the technical component, it is guaranteed that the company's data is assembled from its source systems and that it is stored, organized, and cleansed despite its origin. The business component assures that the structured data allows creating the desired reports and key figures.

The process of extraction, transformation, and loading (ETL) belongs to the technical component and is summarized in three steps. First, extract the data from the source system. Then, transform this data for business level users and, finally, the data is loaded into the final data warehouse tables [10, Chapter 5]. The Figure 2.1 shows an example of an ETL process, where the final result is loaded into the data warehouse and used for reporting or analysis in a data mart.

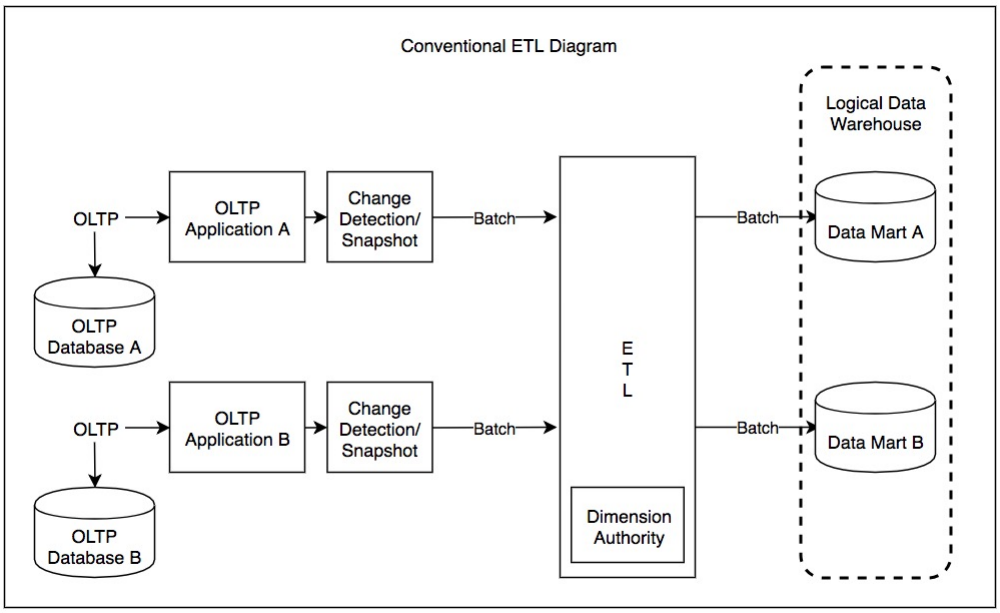


Figure 2.1: Conventional ETL diagram [12].

An alternative process is the extraction, load, and transform (ELT), which is used in systems requiring the support of more continuous, and faster, flow of data into the data warehouse. In this process, the data is extracted, loaded directly into the data warehouse server and the transformation occurs in the data warehousing environment. Finally, the data is loaded into the data warehouse tables [13, Chapter 1].

A data mart derives from the data warehouse and is created with the principal purpose of reporting to a specific group of users and thereby provides easy business access to relevant information. Companies may have several data marts for different departments, for example, human

resources, marketing, and sales. The advantage of that is the shorter time of queries since there is fewer data to process [10, Chapter 1].

Figure 2.2 demonstrates an overview of a data warehouse containing a back room, which stores and structures the data and a front room, based on BI to access the information and make better decisions.

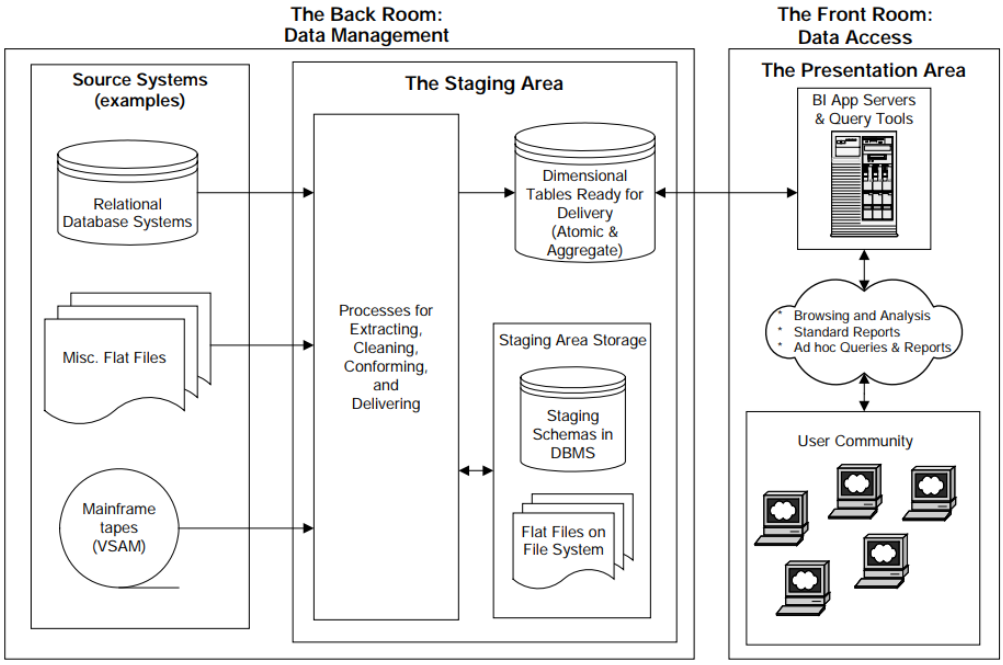


Figure 2.2: The back room and front room of a data warehouse [11].

Laursen and Thorlund stated in [10, Chapter 5] that integrating data into an overall data warehouse as many advantages, such as:

- Avoid overloading of source systems when generating daily reports;
- Integrate data from many different origins;
- Store historical data that can be changed or removed in source systems;
- Add new business logic to data that doesn't exist in the source systems;
- Establish central reporting and analysis environments;
- Secure scalability to guarantee future maintenance of new data volumes;
- Ensure consistency and valid data definitions across business areas.

Focusing on sports organizations, one of the data warehouse's advantages is to collect and organize data in several forms, which allows, for example, to build a fan profile to help and support ticketing strategies.

The customer relationship management (CRM) data warehouse works as a centralized database with integrated information about the customer, namely tickets, food and beverage, and team merchandising items purchased [14].

A practical example could be the analytics department mine the CRM data to identify customers with season ticket that purchase hotdogs or popcorn in the half time of every game, thereby increase stock at those times or advertise coupons to the customers. Another use of the customer profile could be to offer him a beverage or food when the game matches with his birthday.

There is an interesting case applied by the Orlando Magic, which through its data warehouse, found a considerable number of fans that were not buying tickets for multiple games, whether it was in the same season or throughout multiple seasons. The NBA team's research familiar with the gap began working to identify the patterns that led fans to buy tickets. After that, they launched two multi-ticket packages at more affordable prices and the results in the first season were extremely positive [15, Chapter 2].

In 2014, SAS and the New York Mets established a partnership to get the club to use SAS analytics to analyze fan data and customize their experiences on an individual level. By examining, for example, social media data, the Mets can accurately interpret the fans' feelings about a given player or game. With this in mind, through their website or social media channels, they interact with fans using the information supporters care about the most. [16].

2.1.2 Reporting and Dashboard

In BI, reporting and dashboard are ways to select and present information that intends to support decision makers. Throughout this chapter, both techniques are described individually and their related work is shown.

Laursen and Thorlund [10, Chapter 4] define reporting, in the context of Business Analytics, as "selection and presentation of information, which is left to the end user to interpret and act on". They also define it as descriptive statistics, in a statistical context, because the information is merely presented and no exploratory analyses are performed. Reports with descriptive statistics are the most common way to transform business data on business information, as they can be standardized and consequently automated.

Typically, a data warehouse is used together with a reporting solution and rules are defined that guarantee the generation of reports. For example, reports showing monthly sales of game tickets or daily sales of merchandising items, in which the data warehouse feed both reports.

Goldsberry [17] innovated data analysis in basketball when introduced new spatial metrics and advanced visualizations to help the interpretation of the spatial dynamics of NBA teams and players. His research focuses on the shooting performance, in which he presents cases of players who have a high field goal percentage, but who are not seen as the best shooters. He defined the spread and range metrics, which respectively counts how many positions where the player has made at least one shot and the player's shoot effectiveness from the most diverse court locations.

Concerning visual analysis, the research presents heatmaps that demonstrate the spread and range values from different players, which for him the range determines who are the best shooters

in the league. Figure 2.3 shows heatmaps from the four players with the highest range values and it is noticed that each one behaves differently in relation to the position where he prefers to shoot.

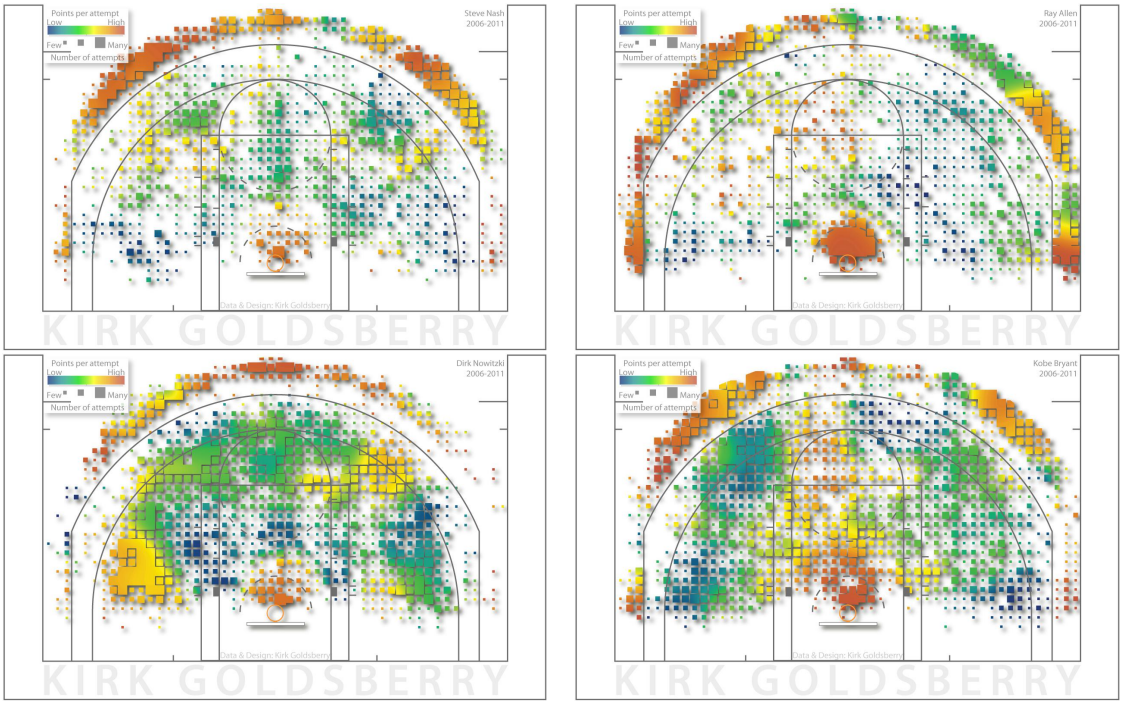


Figure 2.3: The shooting ranges of Steve Nash, Ray Allen, Dirk Nowitzki, and Kobe Bryant [17].

These reports can be made to analyze the shooting trends of the opponent team and based on that the coach might apply a zone defense against his opponent.

Most academic research focuses on developing metrics and key performance indicators (KPI) that could objectively describe a player's or team's effectiveness. However, there is a gap in how sports organizations and athletes make use of this data [18].

To fill this research gap, Caya and Bourdon proposed a value creation framework to describe how and where Business Intelligence & Analytics (BI&A) can generate benefit in competitive sports organizations and athletes.

The framework assigns value creation to three groups: institutional level; organization level; individual level, where all are affected by external actors, who are defined as technology vendors, sponsors, fans, and the community.

The institutional level represents national professional leagues and international sports federations. These macro entities compete against each other to enhance the sports fans attention, which consists of improving and sustaining a public image to attract more sponsors and fans. Seek financial success is also part of the goals. To achieve this can be used interactive data visualization technologies to enrich fans' experience, where the external actors join the purpose to provide the required technology. Television networks and media also help to it, because they own the infrastructure and legal rights to transmit the sports events to fans. Secondly, sports leagues and federations can provide relevant and up-to-date data with statistical analysis through social media

and web portals. The NBA is a good example, as they provide to teams and spectators reports and data representation tools that measure the performance of the players and teams in the league.

Sports teams and their sub-teams, like, management team, coaching staff or support staff define the organizational level. Improve decision making and player preparation is the main focus on this level, but also increase revenues and overall financial performance.

For all this to be possible, the staff needs to shape a mindset predisposed to data-driven decision making. The individual level is composed of athletes, who are the ones converting decisions from the organization level into physical action during sports events. The main goal for them is to have better injury prevention, effective health management, and gather useful insights on how to perform better. Sharing specialized reports and data visualization tools along with coaches and other support staff members allows athletes to identify opponents' patterns of play, receive personalized feedback and learn game tactics. After the game, a player can review his game performance with a personalized post-game video along with specialized statistics.

Lastly, one remarkable example of BI application is how Shane Battier, who played for the Houston Rockets, processed detailed data about his direct opponent's playing patterns before the basketball game to diminish the performance impact of that opponent [19].

About dashboards, Few stated in [20], [21] that dashboards are a visual representation containing sets of performance indicators and key performance indicators, organized and displayed onto a single screen since the goal is to be interpreted quickly.

Unlike reports, dashboards are customisable interfaces that include a combination of text with graphics, and if properly designed they can communicate more efficiently than only text. Secondly, they can be interactive, applying filters the user can pick the information of only one team, instead of all the teams' data. Finally, dashboards can be seen as the first point of information, to keep people aware and take them into a further investigation.

Harrison and Bukstein show in [15, Chapter 13] a framework, developed by them, pointing the main aspects for maximizing the chances that a data visualization communicates effectively (Figure 2.4).

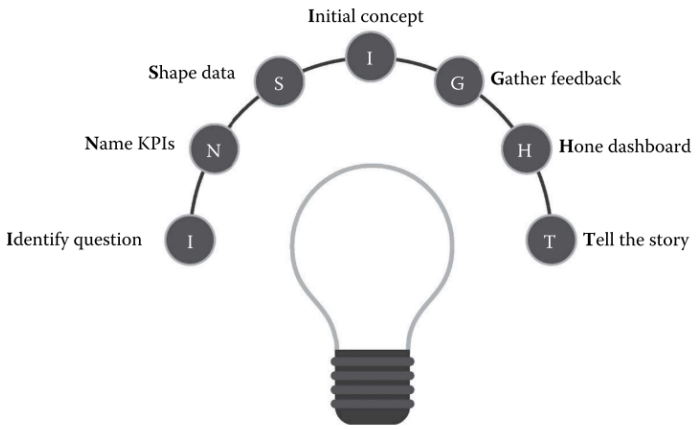


Figure 2.4: Insight framework [15].

In the sports environment, having the most important performance factors available in a simple and accurate way influences any decision-making process. Another fact is that coaches or scouts struggle to remember and recognize important game events [22], [23]. People can see events happening, but only a few details are processed, and not everything that is processed is stored as memories.

Since the data comes from multiple sources, different types of groups want to view the summary information, namely coaches, players, or other management staff. Each group has a different purpose for analyzing the information, so the design and focus of the dashboard must match that purpose [24, Chapter 4].

BKViz [25] is a good example of an interactive dashboard for basketball games. It allows a user to analyze a player's performance, how he interacts with his teammates or even allows to interpret the game style and chemistry of a team. Simple interactions, such as filtering by players on the court or by the game's timeline, enable to interpret how the match developed at a glance (Figure 2.5).

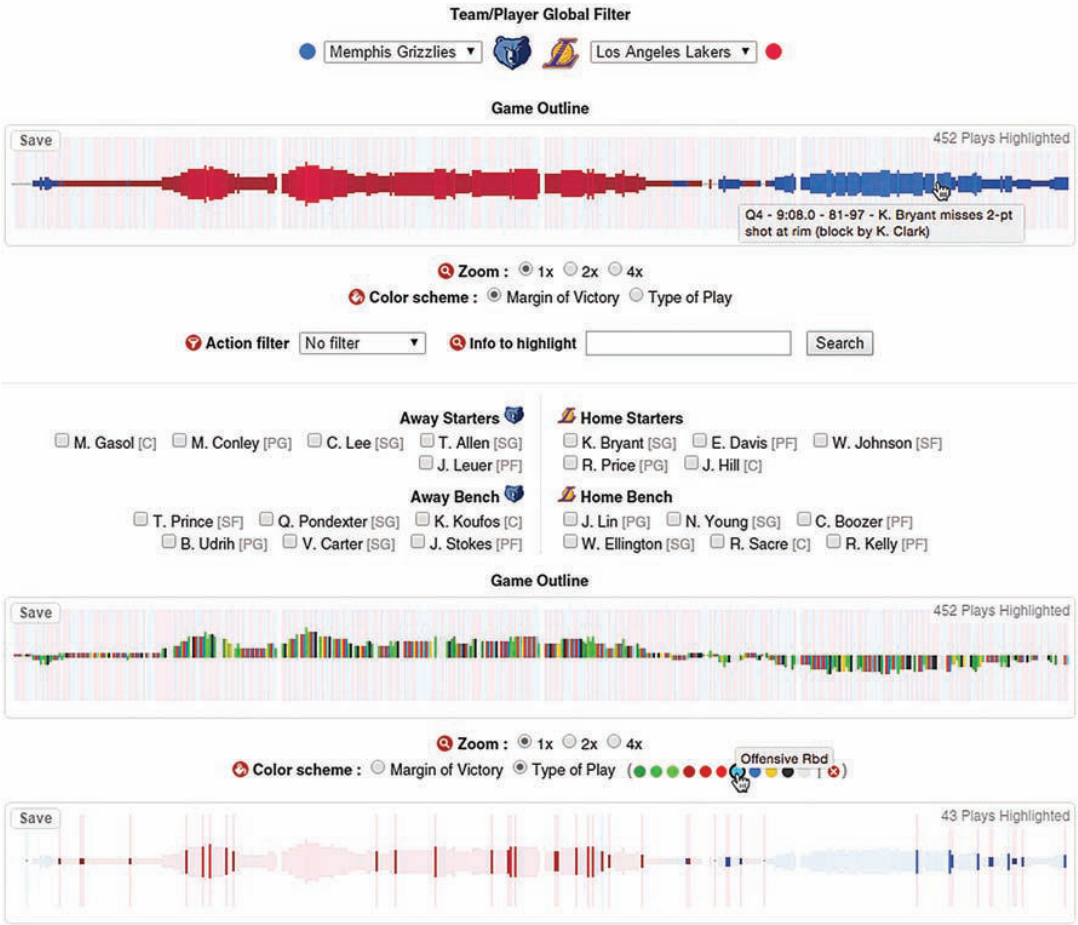


Figure 2.5: BKViz Game Outline feature [25].

In Hockey, Pileggi et al. [26] have developed a system for viewing statistics related to shots,

for example, the location and length. To show these statistics provide three views: shot map, traditional heatmap, and radial heatmap. To give more flexibility to the analysts using the system several filters are provided. Focused on the chemistry between players, when it comes to passing the ball between them, SoccerStories [27] through heatmaps and node-link diagrams demonstrate this relationship to detail.

2.1.3 Existing Products

The next section describes products related to business intelligence on sports.

2.1.3.1 SportRadar Performance and Advanced Widgets

One of the top leaders of the sports, media and betting industries is SportRadar [28], an international organization employing over 2.000 people in more than 30 locations all over the world. They develop many different products related to various sports types, such as widgets for basketball, soccer and American football.

Built for basketball, soccer and National Football League (NFL) the Advanced Widgets [29] mission is to provide better entertainment on the sports fan experience. With this in mind, more specialized information than fixtures, standings, and line-ups is presented through a digital platform.

In the case of basketball and soccer, Figure 2.6 and Figure 2.7 represent a widget's screen, which provides, respectively, a heatmap with the teams shooting zone and league average goals scored and conceded by a team.

Performance Widgets [30] is another product made by SportRadar and focuses on individual or team performance by showing, for instance, 1-on-1 duel performance, passing or shooting accuracy during a game or season. Currently supports soccer and NFL. Figure 2.8 represents a screen that shows the goalkeeper's performance.

2.1.3.2 Dartfish

Since 1999, Dartfish [31] develops products to create, analyze and distribute video content across industries, such as education, healthcare, and sports.

In a range of several products, myDartfish Mobile S [32] and myDartfish Live S [33] bring intelligence to the data producing reports that allow better visualization of the athlete or patient performance and extract detailed knowledge. Figure 2.9 shows a report generated by Business Intelligence tools from myDartfish Live S.

2.1.4 Conclusion

Through the literature, it is seen that BI adds much value in sports, either at the organization level or athlete level. Over the years, professionals are shaping their mindset to be predisposed to use data in any decision-making.

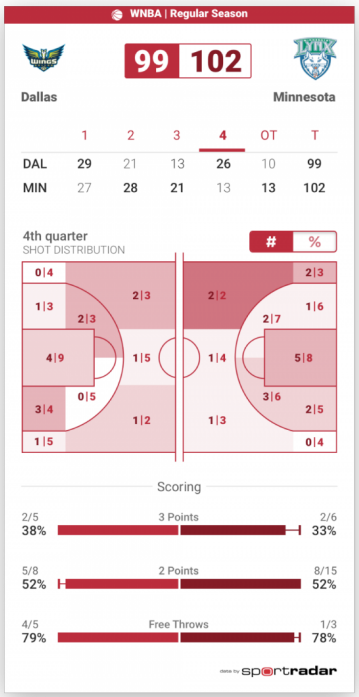


Figure 2.6: SportRadar Advanced Widgets for basketball [29].

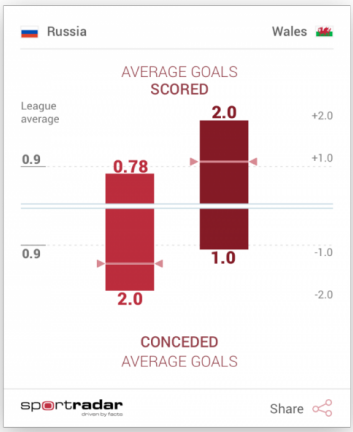


Figure 2.7: SportRadar Advanced Widgets for soccer [29].

In terms of data warehousing, several stories show that clubs with good collect and processing data techniques have innovated and improved their sport. However, no cases were found where clubs integrated more than one sport in their data warehouse. However, no cases were found where clubs integrated more than one sport in their data warehouse.

Both reporting and dashboards are very intuitive and valued tools for the sports staff. Reports are created more easily and can be automated. Dashboards through filters provide more flexibility to the analyst, but the creation process can be more time-consuming.

Since sports analytics are emerging quickly, there are more and more quality products and the market is becoming very competitive.

2.2 Machine Learning

Recently, one of the most popular approaches in data analytics is Machine Learning (ML), which can be used to predict relevant facts. Compared to BI, with this approach, it's possible to do a more advanced analysis, which consists of, per example, train on a data set with known events and use that knowledge to predict the events for another data set.

In Sports, ML allows predicting values, such as, the probability of a player getting injured and what characteristic a team should improve to increase their winning chance.

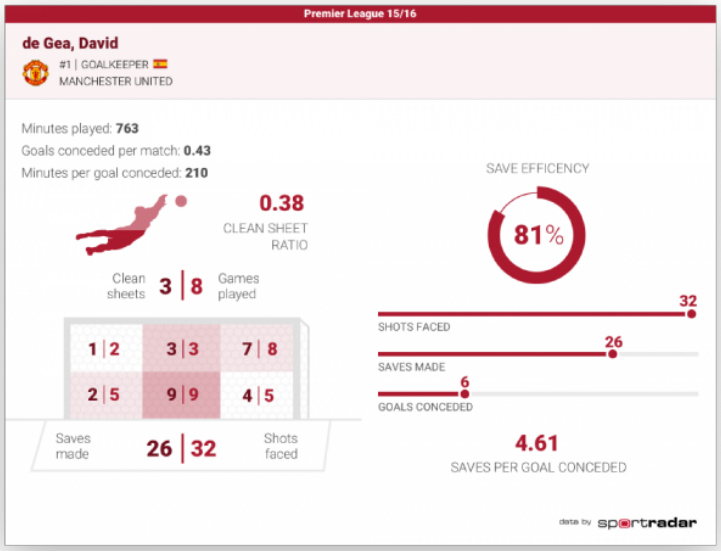


Figure 2.8: SportRadar Performance Widgets for soccer [30].

One of the most influential people, more specifically in baseball, is Nathaniel Silver, better known as Nate Silver. During his career as a baseball analyst, his most notable projects were PECOTA [34] and FiveThirtyEight [35].

PECOTA stands for Player Empirical Comparison and Optimization Test Algorithm and was developed in 2002-2003 focused on player performance prediction for the Major League Baseball (MLB). The algorithm adopts an unusual methodology, instead of centering on player individual statistics, such as batting average or home runs, it focuses on the historical performance of players similar to the player being predicted. Thereby produces a probability distribution of the player's performance over the next five years.

FiveThirtyEight is a website that publishes content related to analysis in politics, economy, and sports. About sports, there are articles pertaining to the NBA, National Collegiate Athletic Association (NCAA) Men's Basketball tournament and MLB.

The following chapters describe ML techniques and examples of applications in the sports environment. This techniques are divided into four categories: supervised learning, unsupervised learning, stochastic model, and Bayesian model.

2.2.1 Supervised Learning

In supervised learning, observations are defined as feature vectors, where each dimension contains the value associated with a given feature. Associated with each one of these observations there is a label, that represents the value the model will try to learn to predict. For instance, if the data collection represents a person, each feature must describe that person. This data may contain

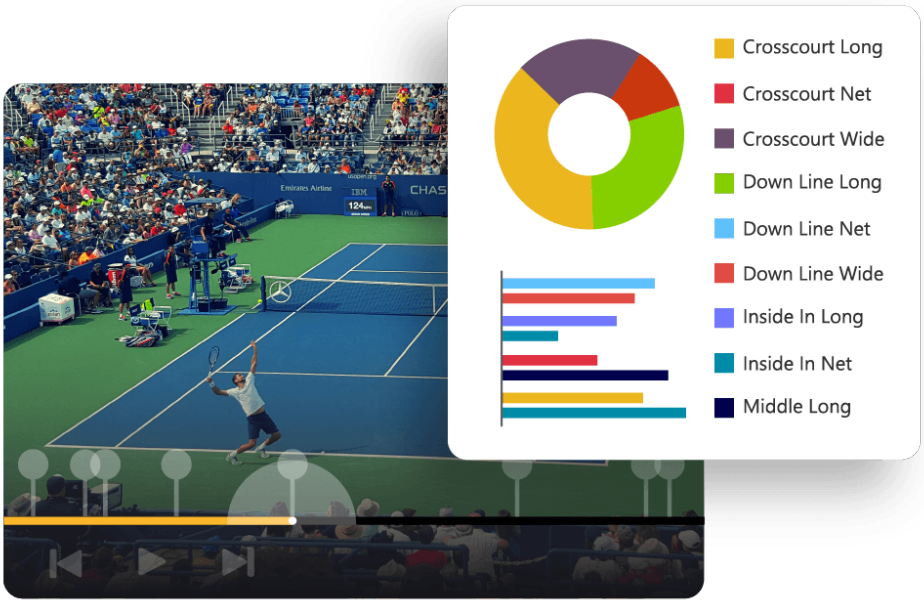


Figure 2.9: BI report example from myDartfish Live S [33].

features such as the person's height or gender [36, Chapter 1]. There are several mechanisms of supervised learning, below are described some of them in a summarized way:

Support Vector Machine

Support Vector Machines (SVMs) are a set of supervised learning methods commonly used for classification, regression, and outliers detection. This technique is effective in high dimensional spaces and also in cases where the number of samples is smaller than the dimension length. Other advantages are memory effectiveness because it uses a subset of training points (support vectors) in the decision function, and versatility since different kernel functions can be specified. On the other hand, SVMs do not give straight probability estimates, it uses an expensive five-fold cross-validation to calculate that [37].

Decision Tree

A decision tree (DT) is defined by [36, Chapter 3] as "an acyclic graph that can be used to make decisions". Each branching node represents a specific feature and if the feature's value is below or above a specific threshold it is followed the left branch or right branch respectively. The decision is made based on the leaf node that is reached. As trees can be viewed it is simpler to understand and interpret the results [37].

Random Forest

Random Forest (RF) is a meta estimator used for classification and regression tasks, which fits multiple DTs on several sub-samples of the data set. When splitting a node of a tree, the split

that is picked is the best split among a random subset of the features. To outputting the class uses the average to improve the predictive accuracy. RFs help to control over-fitting, which sometimes occurs in decision tree [37].

Extremely Randomized Trees

As the name says, in Extremely Randomized Trees, the way the splits are computed reaches an extreme level of randomness. Apart from random forests that look for the most discriminative thresholds, this algorithm draws thresholds at random for each candidate feature, and the best randomly-generated threshold is chosen as the splitting rule [37].

Gradient Tree Boosting

Gradient Tree Boosting is used both in regression and classification problems, which produces a prediction model using decision trees as base learners. The model is trained by iteratively boosting a single tree and after each iteration, the subsequent predictors learn from the mistakes of the previous predictors. As a result, it takes fewer iterations to reach the final predictions. This could lead to overfitting, therefore, it is crucial to choose carefully the stopping criteria [38].

In the sports context, exists many different related works to predict the outcome of a game. Regarding basketball, Pai, ChangLiao and Lin [39] developed a hybrid model integrating Support Vector Machine technique and a decision tree approach, called HSVMDT, to predict the outcome of a game. Furthermore, the rules generated from the HSVMDT model enables coaches to learn what factors can increase the chances of winning a game. In the Chenjie Cao's dissertation [40] demonstrates a data analysis of six NBA seasons using Simple Logistics Classifier, Naive Bayes, Support Vector Machine, and Artifical Neural Networks. The first 5 seasons were used to model fitting and the 6th season was used to the score purpose, being that Simple Logistics showed a better prediction accuracy comparing to the others. Also about the NBA, Valenzuela [41] used ensemble learning methods to improve the accuracy of single models, such as SVM and random forest. In addition to the NBA, Zimmermann [42] applied various supervised learning techniques to predict the National Collegiate Athletic Association Basketball (NCAAB) seasons and makes a comparison between the predictions obtained in NBA and NCAAB. The author concludes that some metrics used in the college matches are easily applied in the NBA and that turns out to be some of the most relevant data. However, at the level of each individual game, he thinks the NBA is harder to predict.

For soccer, Danisik, Lako, and Farkas [43] developed an LSTM neural network focusing on the player attributes to predict outcomes. They use the player attributes from the video game FIFA among with match history from the five best European leagues. Ulmer and Fernandez [44] applied five different classifiers using a training set composed of ten EPL seasons, and a test set composed of two EPL seasons. The features focused on whether a team is playing a home or away game, and the team form until the game.

2.2.2 Unsupervised Learning

Unsupervised learning deals with extracting patterns from sets of unlabeled data. This however means that the usual evaluation metrics that are used in supervised learning are unavailable. In this approach, the goal is to build a model capable of using a feature vector as input and transforms it into another vector or into a value that could solve a practical problem. Two tasks of unsupervised learning can be clustering the data into groups by similarity and dimensionality reduction to remove redundant or highly correlated features [36, Chapter 1]. These tasks can be applied, for example, to divide the population by location and purchasing habits in order to provide useful publicity. Or, a data science team reduces the dimension length in a large data set to apply simpler algorithms and decrease file size [45].

Below are described two algorithms for clustering the data in a summarized way:

K-means Clustering

In k-means clustering the goal is to cluster the data points into K groups. To define the groups, it's assigned a centroid (the core of the cluster) for each group. The algorithm's first step is to randomly choose the centroids. Secondly, for each data point find the nearest centroid, often using the Euclidian distance, and assign that centroid to the data point. Thirdly, calculate the new position of each centroid applying the average position of all cluster's points. The last two steps are repeated until there are no more changes in the centroids [45].

Hierarchical Clustering

In hierarchical clustering beyond cluster assignments, it builds a hierarchy between the clusters. Each data point is assigned with a cluster, resulting in N clusters. The further step is to merge the two clusters that are closest to each other. The third step consists of recomputing the distances between the clusters. One way to do this is to consider the distance between two clusters to be the average distance between all their respective members, which is called average-linkage clustering. Finally, the last two steps are repeated until one cluster of N data points remains, which turns out to be a tree [45].

For dimensionality reduction the following algorithms can be applied:

Principal Component Analysis

Principal Component Analysis (PCA) provides dimensionality reduction while maintaining structure (variance). PCA remaps the space in order to make the transformed dimension smaller than the original and to better comprehend high-dimensional data [45].

Singular-Value Decomposition

Singular Value Decomposition (SVD) is a matrix decomposition method for decreasing a big matrix into a product of three smaller matrices. It is commonly used in data pre-processing, which

enables, for example, to reduce the values of an image matrix without significant loss in image quality [45].

In Tennis, unsupervised learning methods are applied to identify types of serves directly from data. In the work of Wei et al. [46] it's applied clustering methods in Hawk-Eye data from 3 years of the Australian Open Men's. Through the identification of serves, they have been able to predict what kind of serve a specific player has more tendency to do, considering the game state.

2.2.3 Stochastic Model

Stochastic modeling is a form of modeling usually defined as a collection of random variables. According to Kenton [47], the goal of such modeling is to "estimate how probable outcomes are within a forecast to predict conditions for different situations".

An example of a stochastic model is a Markov model, which depends only on the current state to determine future states. This is known as the Markov property, which states that given the present state, the next state is only dependent on the current state [48, Chapter 1]. There are four types of Markov models, which are used depending on the situation. When all states are observable Markov chain or Markov decision process (MDP) is used. When the system contains states that are not observable it uses the Hidden Markov model (HMM) or Partially observable Markov decision process (POMDP).

In MIT Sloan Sports Analytics Conference 2014, Cervone et al. [49] proposed a framework to assign a value to each moment of possession in a basketball game, called expected possession value (EPV) based on how many points the offense is expected to score. The possession model is supported in a Markovian assumption and allows to estimate both the probability of a player to take a specific decision in a given situation and the corresponding EPV of the possession when the player makes that decision. This framework uses player-tracking data from the NBA season 2012-13 and helps to answer questions about players, such as, if he makes better decisions than the league-average player or if he is a selfish shooter. A common technique for modeling a version of the sports field is to apply a Markov chain, which is a stochastic process, as Goldner applied for football [50]. His work also establishes the expected scored points.

Other models can be used to determine the likelihood of a match result in which other variables are taken into account. In soccer, the work [51] applies the Bradley-Terry model, supported by team factors, such as if the team was recently promoted to the league, the number of days that rested before the game and its current form. Also, they focus on the hierarchical Poisson log-linear model to directly predict the goals scored by each team in each match.

Coefficients representing a specific tactic, whether in attack or defense, are significant, as some teams may perform better or worse depending on the tactic they are using. Identifying what conference the team belongs to, in the case of the NCAA basketball tournament, enables using the general behavior about the attack or defense of that conference. These coefficients can be used in Poisson distributions [52].

2.2.4 Bayesian Model

The Bayesian model belongs to the group of statistical modeling that is one of the innumerable applications of Bayesian methods. These methods use Bayes' theorem to compute and update the probabilities after collecting new data. The Bayes' theorem determines the conditional probability of an event, based on previous information or conditions related to it.

Given two events A and B, the conditional probability of A knowing that B is true is expressed as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

In sports, a new method was applied using Hybrid Bayesian networks [53], which belongs to the group of Bayesian models, that predicted results of soccer matches from leagues around the world. That is, instead of using data from only one soccer league, it uses several leagues, such as the Japanese and Russian championship, to improve the training data.

To make this possible three aspects are taken into account: temporal data, new team data, and different leagues. The temporal data means that considering games between team X and Y in season 2016-2017, the previous matches between them become less relevant as they are further away from the season 2016-2017. Sometimes two teams have never met each other since teams are promoted or relegated from a league every year, which results in no history of results between them. To handle this problem, the new teams are assigned with a default rating of 0. Since a team rating in league A is different from a team with the same rating in league B, different league games are used only when the difference rating of these matches exhibits strong similarities.

2.2.5 Kaggle Competitions

There are several products on the market relating to odds in sports, namely the odds of a team winning a game. However, companies do not share how these probabilities are determined, so it is not possible to know if they apply machine learning techniques. Thus, in this section, rather than referring to the existing products, it is talked about the Kaggle competitions.

Kaggle competitions are a good initiative that brings competition for the best prediction model and enables the sharing of knowledge in several domains where prediction is relevant. Kaggle [54] is an online community for data scientists, which provides users to publish or find data sets of several different themes and also enter competitions to solve data science challenges. These competitions aside from rewarding the scientists who built better models and consequently had better predictions, also, encourages the exchange of knowledge, where often the authors of the best works explain their ideas and share their code. Regarding sports, a good example of a basketball competition is the March Machine Learning Mania, which had contests from 2014 until 2017. This competition rewards the best forecast for the NCAA bracket of March Madness.

Zach Bradshaw [55], who won the 2015 contest, used the previous experience at modeling NBA games from his previous work as a sports analytics expert at ESPN. This experience helped on the data pre-processing and the modeling techniques applied in the competition. He applied a Bayesian framework to incorporate prior knowledge that was not on the contest's data set. However, he manually tweaked a prediction for one game, which resulted in a successful prediction of an upset.

In 2016, the winner was Miguel Alomar [56], which applied a logarithmic regression and random forest. The model key factors were the offensive and defensive efficiency. He also updated the team weight according to the strength of schedule, therefore penalizing teams that did not play against the best teams in the championship.

2.2.6 Conclusion

The state of art shows much quality and as it is a popular and growing subject, these works use recent data sets.

Currently, there is a lack of approaches based on the different combinations of lineups, which is a very important factor in predicting or simulating a game. Moreover, some work only uses one analysis mechanism which doesn't allow to compare the efficiency of the outcome with other mechanisms.

In short, there is still a lot of methods to explore and the wide range of mechanisms also helps in this exploration.

As mentioned previously, in this section no existing products were described, however, during the research it was found that the products in the market do not focus on the combination of lineups, similarly to the related work. The target audience turns out to be more directed to fans, who use the team winning probability as a support for their bets, or discussions with their friends. These odds do not give useful insights to coaches because they can not explore or see the lineups that led to these predictions.

In conclusion, there is a lack of products that support the coach decision concerning his lineup, that allows him to analyse different situations taking into account the opponent and the game style he intends to use.

2.3 Simulation

Apart from the other techniques, Simulation allows seeing the intermediate results that lead to the outcome. Computer simulation consists of an attempt to model a real-life situation in order to understand how the system works. Basically, it tries to predict the system's behavior from a set of parameters and initial conditions. An overview of this process is represented by the following picture:

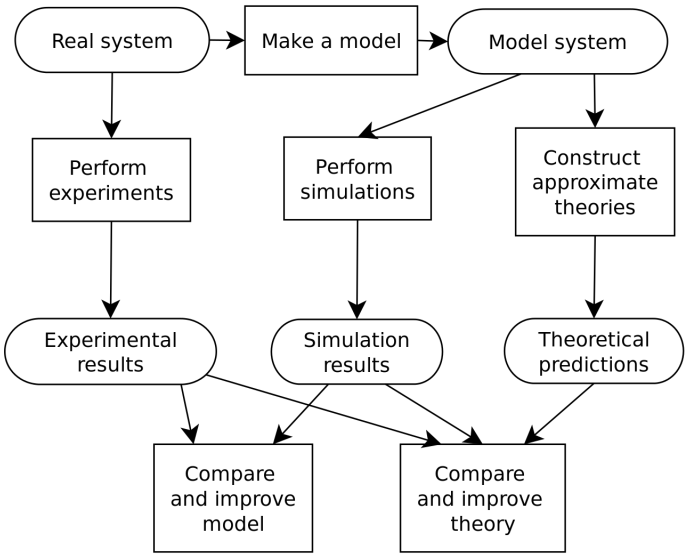


Figure 2.10: Process of building a computer model, and the interplay between experiment, simulation, and theory [57].

Oh et al. [58] introduced a graphical model for basketball match simulation in MIT Sloan Sports Analytics Conference of 2015, which the major goal is to bridge the gap between player identity and team level network. The developed model is calibrated with data from the NBA season 2013-2014 like player tracking, play-by-play game log, and line-ups. So, the model encompasses behavior about every touch and event in a game as a progression of transitions between states. This approach consists in calculating the probabilities of shot frequency, shot efficiency, pass between two players, shooting foul and corresponding free throw, rebound and turnover. Therefore, this work helps answer questions regarding the team, for instance, what team will win a match, or what teams will be in the season's top 3, but also questions regarding the player selection, such as how well a player will perform against a specific opponent.

Moreover, Vračar et al. [59] uses a Markov chain to simulate play-by-play in basketball with similar events but taking into account situational variables. This means considering the difference of points between two teams and how much time is left to finish the game because teams play differently when the game is tied or when the point difference is greater than 15. Also, the same score at the half-time or 2 minutes from the end of the game has a completely different meaning. To avoid invalid states due to game rules, they applied a decision tree and a separate regression model.

Currently, the simulation models of the literature do not allow coaches to enter parameters, like high press on the field. These parameters are useful because if a coach knows that the opposing team plays poorly against the high press, they can simulate the game using this parameter to get an insight about what players are the most capable of pressing high on the field.

Chapter 3

Multimedia Augmentation in Sports

In sports, multimedia applies to a diversity of topics, such as improving a spectator's experience and help players or coaches analyzing games.

Concerning the spectator, it can take the television experience to a whole different level, whether it is before, during or after a game. It allows an offside, in case of soccer, or how a team defends a ball screen, in case of basketball, to be viewed and analyzed in detail. Typically, this improvement consists of overlays over the main content.

On the other hand, extracting knowledge from games and training videos allows to detect the team's weaknesses, to prevent injuries and to increase performance [60]. For instance, record a player shooting a ball enables him to detect and correct small details in his technique that he had never noticed before [61].

3.1 Video Analysis

Video analysis in sports is increasingly important and has grown fast with technological improvement, whether by improving video record or improving player tracking equipment. Work in this area is widespread, although it focuses on the more popular sports. According to Shih's survey [62], more than 80% of the articles are dedicated to baseball, basketball, soccer, and tennis.

The survey classifies techniques in video analysis in terms of semantic level, for instance, highlights detection and event or object recognition. Fig 3.1 represents such classification.

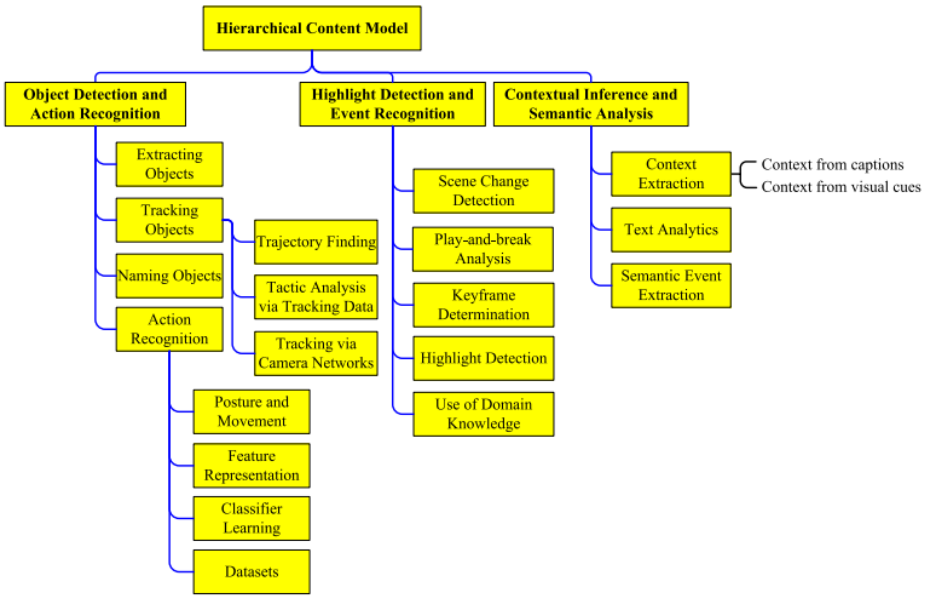


Figure 3.1: Classification of research with different applications according to the semantic level [62]

Detecting patterns in the team defense or attack, recognizing their formation on the field, or identifying the zones in which a team focuses more or less are key aspects that tactical analysis brings. At the individual level, video allows a detailed analysis in the athletes' moves, which it is useful to identify weaknesses in their performance.

About tactics recognition, the work [63] developed a system to detect screen-strategy in a basketball game. A screen is the action of a player blocking the defender's movement by standing next to or behind him. In addition to identifying screens, the system also enables distinguishing screen types. To help professionals identifying the team's offensive trends, Chen, Su, and Hsiao [64] through a broadcast basketball video reconstruct the players' trajectory. These trajectories are presented on a standard basketball field model by a homography transformation, as can be seen in Figure 3.2.

Furthermore, Hobbs et al. [65] using z-score heat maps describes the effectiveness per court location of each team in the women's basketball tournament at the 2016 Olympics Games. It also establishes the sequence of ball-movement performed more often.

On individual performance, coaches can monitor the trajectory of the ball and the parameters during the execution of a free throw, which helps to correct the athlete's skill. Consequently, they can provide feedback concerning the player's body posture or the throw's strength to increase the shot's efficiency [66].

Nowadays, extracting relevant information such as the players in the field or the game actions that took place is a plus. Because of time-consuming and tedious processes, more and more research is emerging to make these methods more effective, efficient, and faster.

There are multiple cases, either in soccer [67] or football [68], where, for example, through

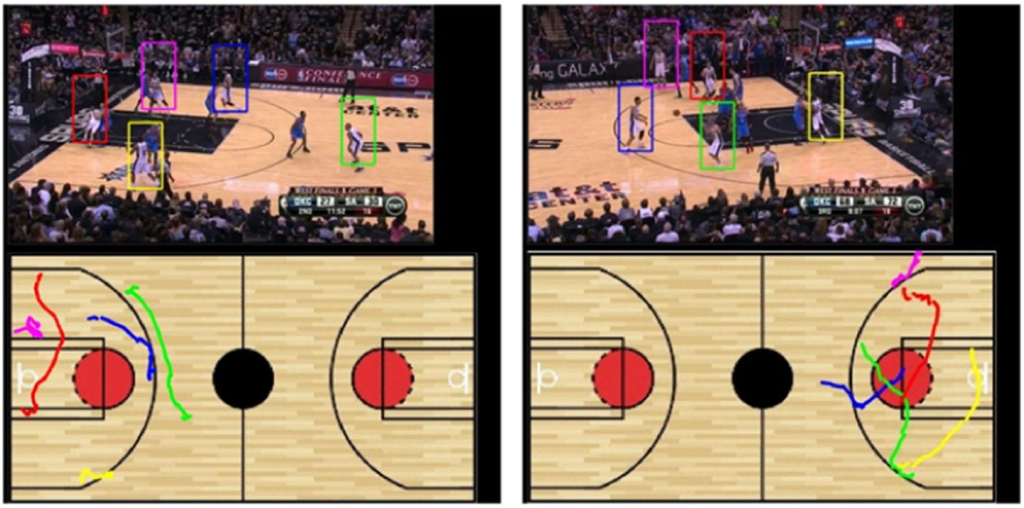


Figure 3.2: Players trajectories are shown on the court model [64].

player recognition, it is identified the video frame in which the teams line up and the respective line of scrimmage for that football play.

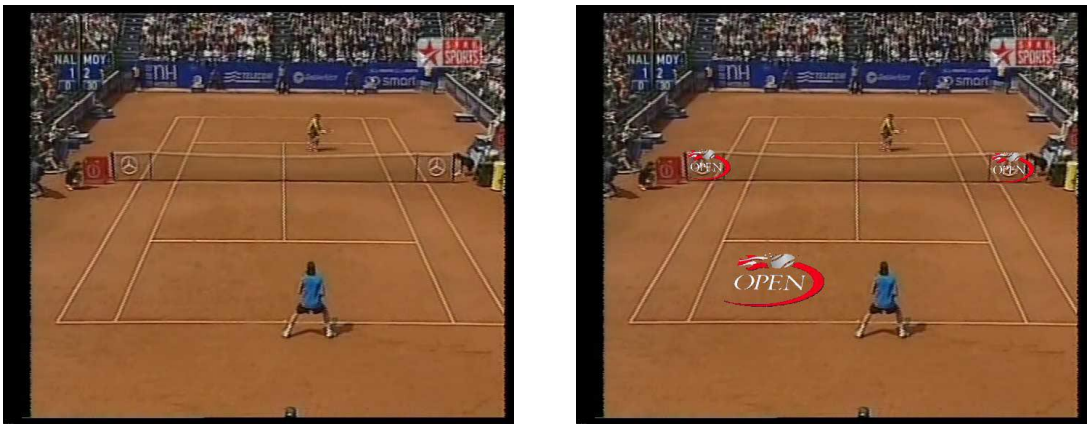
A different method is presented in [69] that first divides the video frames into two classes, with and without a scoreboard. Then, extracts semantic events from the frames classified with scoreboard and extracts replays of the frames without a scoreboard. Through the scoreboard layout of the broadcast video, Yu and Ding [70] detect the interruptions of a basketball game by reading the two clocks presented in the layout.

3.2 Virtual Content Insertion

Recently, the purpose of inserting advertisement and video augmentation has been expanding in sports. Virtual content insertion (VCI) is an application of video analysis and has three important keys: when it is triggered, where it is displayed and how it is inserted into a video [71].

For the virtual content injected to sound realistic, it is crucial that the system acquires the accurate camera matrix for each frame of the video. In the article [72], a new algorithm is introduced for the 3D camera calibration of a broadcast tennis video. As a result, they present a system to insert virtual content correctly, as shown in the Figure 3.3. For the same purpose, but based on visual attention analysis, Huiying et al. [71] developed a system to perform dynamic insertion of virtual content. The time and position of the content are selected through temporal and spatial attention analysis. Another work is [73] that introduced an algorithm with added robustness against unwanted camera movements.

About advertising, existing methods for baseball [74] and soccer [75], automatically insert advertisement without interfering with the main content of the broadcast video.



(a) Original frame. (b) Frame with projected logo.

Figure 3.3: Frame with projected logo [72].

3.3 Existing Products

There is a wide variety of products on the market that support sports content streaming entities in enhancing the viewer's experience or helping coaches with some insights from video analysis.

3.3.1 Dartfish

As seen in Chapter 2.1.3.2, Dartfish provides intelligence to data through its products, but it also has products that enhance the visualization of an action made by a team, an athlete, or a patient. MyDartfishPro S [76] is an example of this since it allows to illustrate the video content with arrows to show the player movement or spotlights to grab attention to a specific player (Figure 3.4).

On the other hand, designed for biomechanical analysis of sports performance, myDartfish360 identifies and corrects technical movements. Through slow-motion playback and taking measurements of angles and distances of the player's movements (Figure 3.5).

3.3.2 Vizrt

Vizrt leads the media and entertainment industry through sports analysis, real-time 3D graphics, studio automation, and management tools. With 40 offices located around the world, they manufacture products used by the world's leading media companies, such as Fox Sports, Sky Sports, CNN, among others [76].

Vizrt's 3D analysis tools give viewers a better understanding of what is happening during a game. These tools allow broadcasters to create a variety of augmented reality graphics, for instance, heatmaps, replay graphics to review critical referee calls, distance lines, and player lineups (Figure 3.6). An example of the application of these graphics is the quick in-game replays.

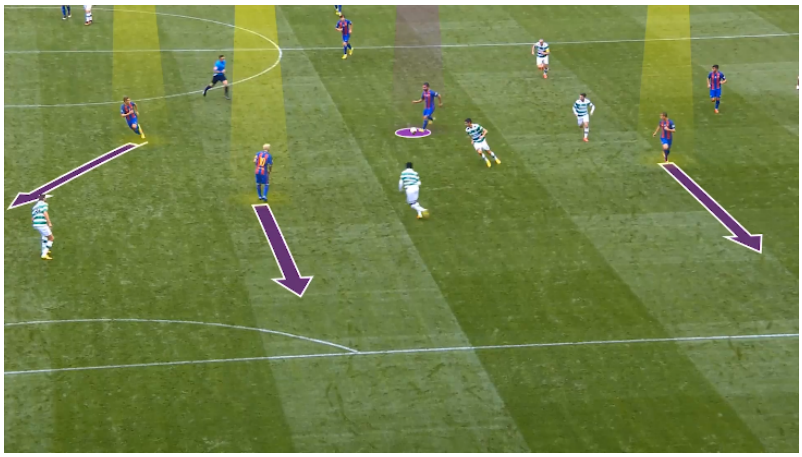


Figure 3.4: Soccer match with illustrated graphics [76].

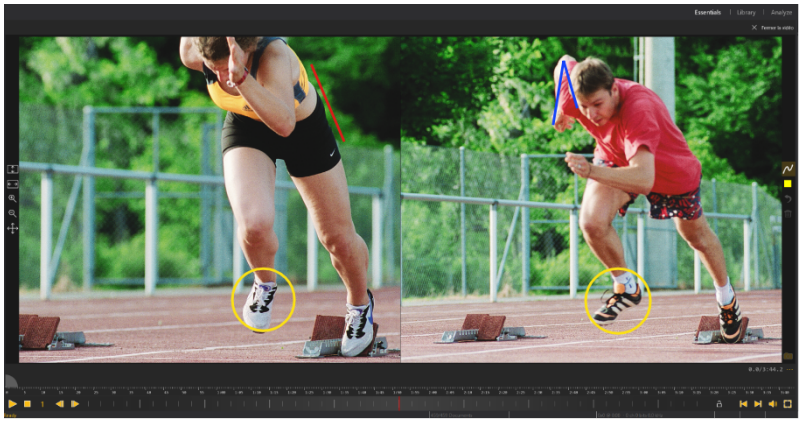


Figure 3.5: Example of player's movement analysis using myDartfish360 [77].

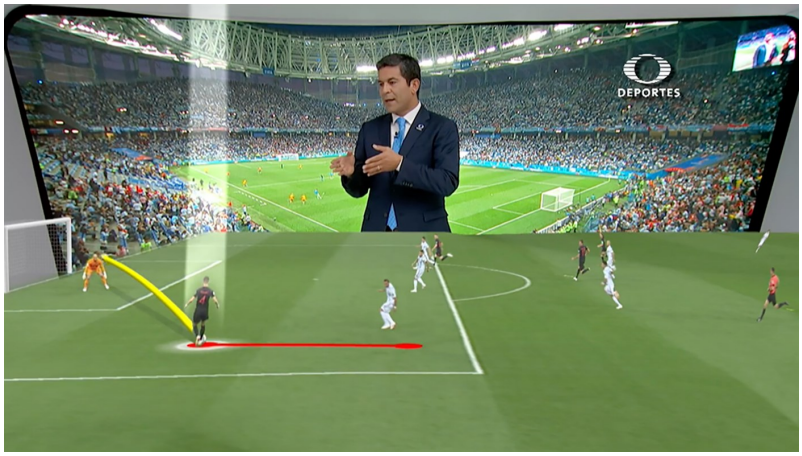


Figure 3.6: Application example of Vizrt's 3D analysis tools [78].

3.4 Conclusion

Broadcasters' tools for sports continue to grow and the competition is huge, either by improving existing features or by developing new ones.

As for the spectators, the increase of the content during a game depends a lot on the information and how it is presented.

There is still room for innovation as data on sports keeps increasing, whether through statistics or players' tracking.

Chapter 4

Data Warehousing

This section details the process of creating a data warehouse. The data warehouse goal is to integrate basketball and soccer data, more specifically the NBA and the EPL.

The following figure demonstrates the architecture defined for it:

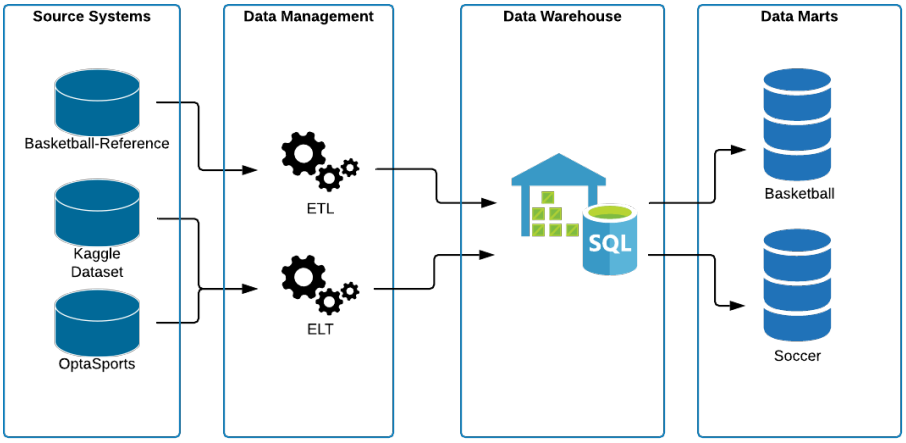


Figure 4.1: Data warehouse architecture.

For each sport is explained bellow which are the source systems, the processes of data collection and transformation, and the final data mart.

4.1 Basketball

This section describes the data sources used for basketball, as well as the extracted data, and the extract, transform, and load (ETL) process that was applied.

4.1.1 Data Sources

The data extracted belongs to the Basketball-Reference website [79], which was launched in 2004 and belongs to the company Sports Reference LLC. The company began in 2000 to provide only baseball information and has now expanded to basketball, soccer, football, and hockey. For the purpose of this study were used the seasons of 2012-2013 until 2015-2016 of the National Basketball Association. The data consist of information about the games that have taken place, the statistics of the players and teams for each game and the details of each team and player.

Figure 4.2 shows information of a few games from October displayed on the website. It's possible to collect the game's date, the home and away team, the game result and the attendance.

Date	Start (ET)	Visitor/Neutral	PTS	Home/Neutral	PTS		Attend.	Notes
Tue, Oct 30, 2012	7:00p	Washington Wizards	84	Cleveland Cavaliers	94	Box Score	20,562	
Tue, Oct 30, 2012	10:30p	Dallas Mavericks	99	Los Angeles Lakers	91	Box Score	18,997	
Tue, Oct 30, 2012	8:00p	Boston Celtics	107	Miami Heat	120	Box Score	20,296	
Wed, Oct 31, 2012	8:00p	Sacramento Kings	87	Chicago Bulls	93	Box Score	21,313	
Wed, Oct 31, 2012	7:30p	Houston Rockets	105	Detroit Pistons	96	Box Score	16,646	

Figure 4.2: A fraction of NBA games information from Basketball-Reference website [79].

Concerning the game's statistics, basic and advanced statistics are both displayed. The basic statistics are, for example, minutes played, field goals, and three-point field goals. The advanced statistics are calculated through basic statistics and are, for example, true shooting percentage, offensive rebound percentage, and defensive rating (Figure 4.3).

The data for teams consists of basic information (name and location), staff and players information, year-by-year stats, and the best players from the club history. For each player presents positions where he plays, his birth, relatives, titles achieved, and several stats, for example, career summary and playoffs play-by-play.

4.1.2 Extract, Transform and Load

For data management, an ETL process (described in section 2.1.1) was adopted using the Python [80] programming language.

To extract the data from the website, a web scraping script was implemented working with the Basketball Reference Web Scrapper [81] and Beautiful Soup [82] libraries. Web scraping involves fetching the web page and extracting the data from it.

The Basketball Reference Web Scrapper provides functions that extract the game calendar of an entire season and the basic statistics of the players and teams per day. All of these functions allow saving the results into CSV files. The data extraction process starts by collecting the game days for each season. Afterwards the basic statistics of each player and their teams is obtained from the games list. The scrapping process is time-consuming so it was decided to use the filesystem as a temporary storage for the CSV files, for each season it generates three files corresponding to the games information, box scores of the players, and box scores of the teams. Subsequently, the data was accessed more easily and without waiting time.

	Basic Box Score Stats																	
Starters	MP	FG	FGA	FG%	3P	3PA	3P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	+/-
Elton Brand	36:12	3	10	.300	0	0		2	2	1.000	2	9	11	2	1	1	0	4	8	+7
Shawn Marion	32:53	5	11	.455	0	0		1	1	1.000	2	7	9	4	1	1	2	4	11	-2

	Advanced Box Score Stats														
Starters	MP	TS%	eFG%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	ORtg	DRtg
Elton Brand	36:12	.368	.300	.000	.200	6.6	25.9	17.0	7.4	1.5	2.1	0.0	13.9	97	96
Shawn Marion	32:53	.481	.455	.000	.091	7.3	22.2	15.3	17.9	1.6	2.3	14.9	18.9	100	96

Figure 4.3: Example of a game's box score from Basketball-Reference website [79].

The second package is a library that allows to navigate and search an HTML parse tree. It was only used to extract the players' position because the previous package does not support players' information extraction. This field can have more than one value, which corresponds to the positions where the athlete played more times during a season.

The Basketball Reference Web Scrapper doesn't have the ability to extract advanced statistics. In order to obtain this data an extension was implemented. This extension follows the same structure and functionality of the basic statistics functions already provided in the package, with the purpose of contributing this code later to the package repository.

There was also a need to limit the web scrapping process in order to prevent the source system from being overloaded. The library ratelimit [83] was used to limit the number of times a request can be executed during a time interval.

About the transform and load processes, Figure 4.4 gives the flow's overview and the resources needed for each step performed.

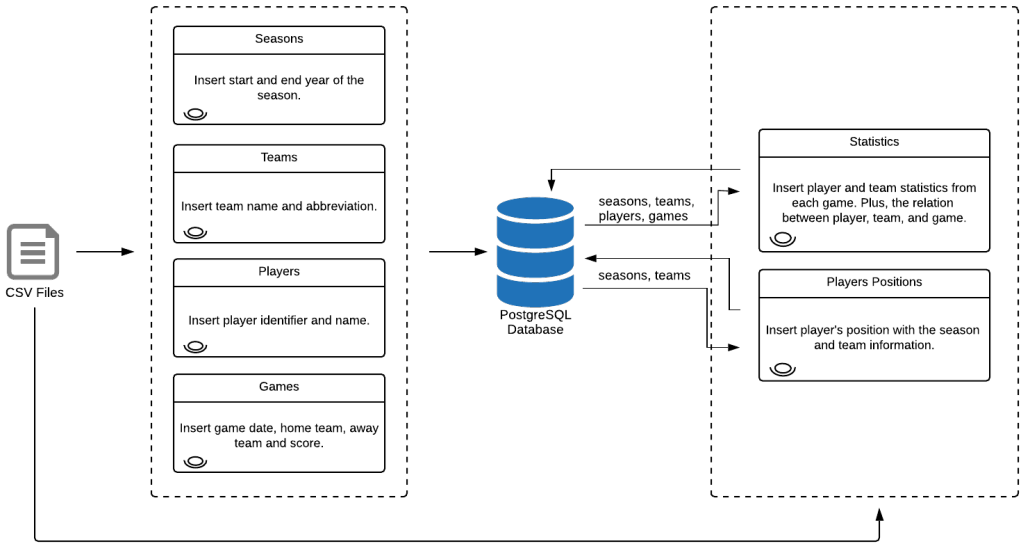


Figure 4.4: Transform and load processes applied in basketball.

The transform and load phases start by creating a schema with the entities tables to execute on the PostgreSQL [84] database. After this initialization the data that has no dependencies is loaded

into tables, namely the seasons, players, teams, and games. Finally all the statistics for players and teams are linked with each game, together with the positions in which the players have played across seasons.

To help out the process, the libraries Pandas [85] and Psycopg 2 [86] were respectively used to efficiently read the CSV files and to inject the data into the database.

Throughout the implementation, one problem arose because some NBA teams over time have changed their name. What happens is that the game data of these teams according to the chronological line may have a different team name. It is necessary to keep the name's background, but also to group the entire history of a team regardless of their name at the time the match happened. Therefore, this information was standardized in a custom table, in which the team's current id, current name and old names are stored.

4.1.3 Data Mart

Concluded the ETL process, the final data mart was defined as follows:

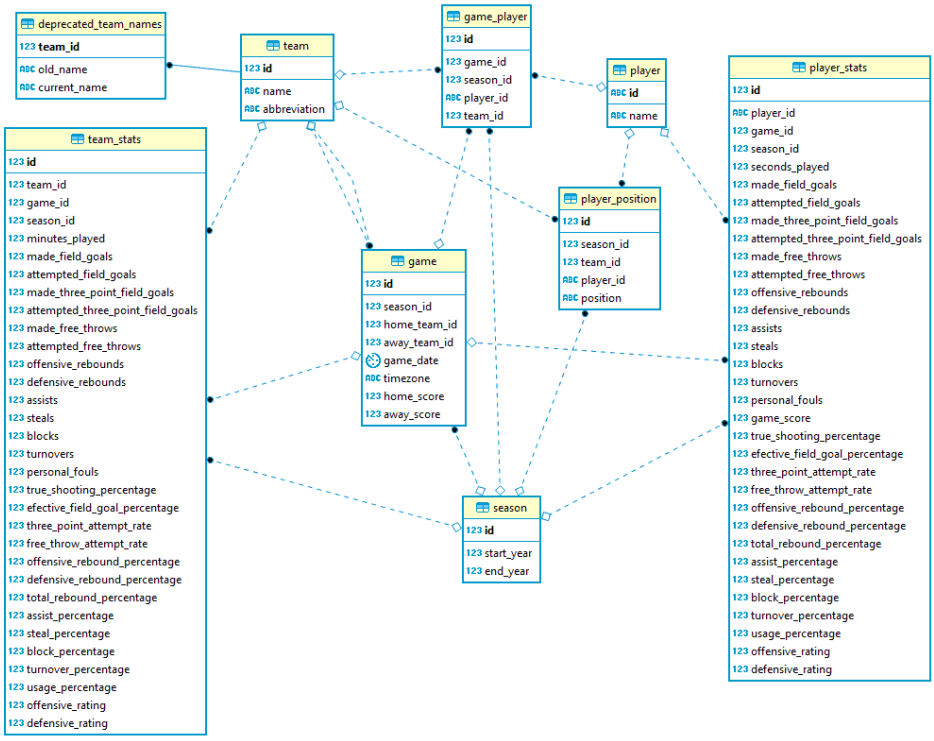


Figure 4.5: Entity relationship diagram of basketball data mart.

4.2 Soccer

This section describes the data sources used for soccer, as well as the extracted data, and the extract, load, and transform (ELT) process that was applied.

4.2.1 Data Sources

The data extracted belongs to the OptaSports [87] and Kaggle. OptaSports is one of the world leaders in sports data and since 2013 is the official data partner of the English Premier League. Kaggle is an online community for data scientists, as previously mentioned (Section 2.2.5).

The data from OptaSports used consists of information about the games, like the game events and teams' formation, about the teams, such as name, stadium, and location, and about the players, namely name, birthday, among others. From Kaggle, it was selected the data set from the user ShubhamPawar [88] that has the players and teams statistics from some EPL seasons. For the purpose of this study were used the seasons of 2014-2015 until 2017-2018 of the English Premier League.

4.2.2 Extract, Load and Transform

Different from Basketball, an ELT process (described in section 2.1.1) was adopted for the soccer data management. The main difference is in the code environment since only a few lines of Python language were implemented and the core process is all performed in the PostgreSQL environment.

Beginning with extraction, since the Kaggle's data set holds the entire information in JSON files, all requests made to the OptaSports API were also stored in JSON files. In this way, data is backed up into JSON files in order to be accessed quickly and without waiting time. As PostgreSQL supports JSON storage directly to the database, the next step was to inject all the data into staging tables through Python scripts. Before the data insertion task, a small organization of the data was performed so that later, within the database environment, access to JSON would be simpler. That is, the data was separated into staging tables, for example, a table for teams information and another for games information. Therefore, as a final result of a staging table, it would be, for example, a column with the team id and another column with the entire team information structured in JSON format. Figure 4.6 shows the database staging tables.

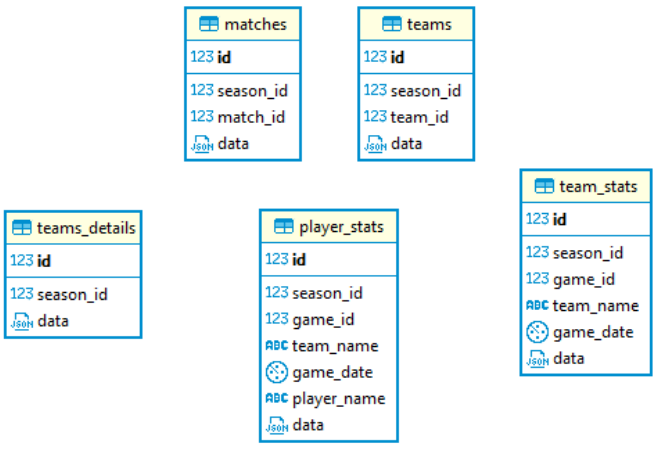


Figure 4.6: Soccer database staging tables.

After the load phase, there is a process of transforming the JSON data from the staging tables to information organized in final relational tables. Figure 4.7 summarizes the data inserted from the transformation result.

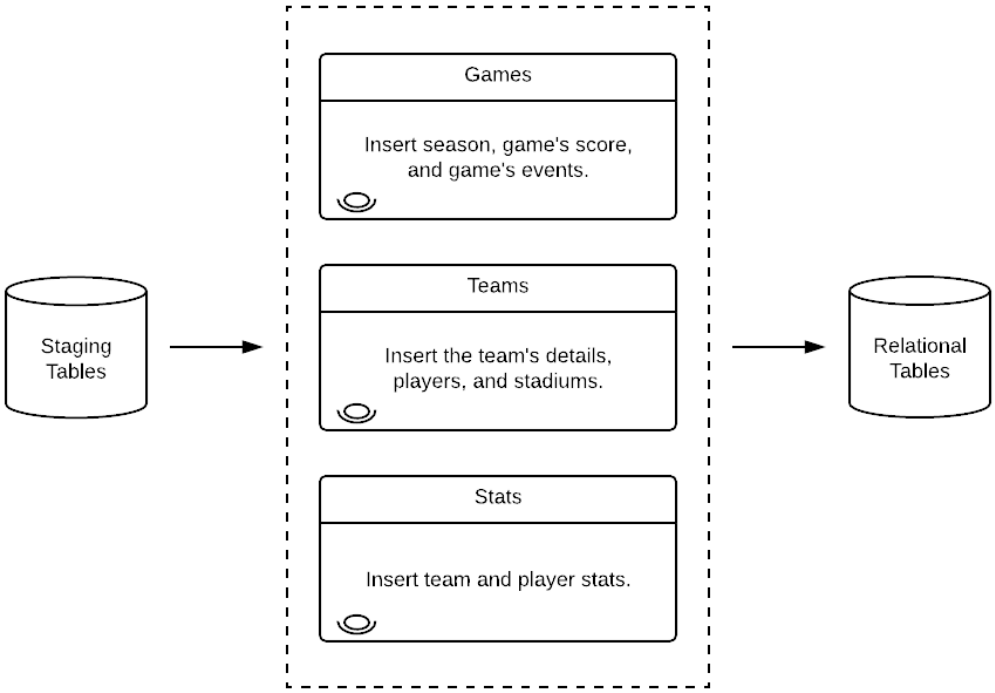


Figure 4.7: Soccer database transform process.

In addition to filtering JSON to get the desired fields, there was also a need to integrate both data sets. Each data set had its own identifiers, so it was required to merge those identifiers into one. We opted to use the OptaSports identifiers and therefore, from that moment all players and teams from the Kaggle data set had to match with the other identifiers. One problem encountered was the names of some teams and players in one data that were not exactly the same as the other data. An approximate search function was applied, which solved all the cases found. Apart from filtering and combining data sets, the data was cleansed because there were null, duplicate, and wrong values.

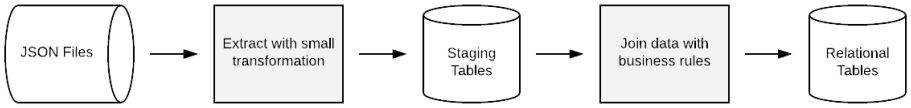


Figure 4.8: Soccer ELT process overview.

Clearly, in this case, it was an advantage applying the ELT process (Figure 4.8), because combining the data sets and cleansing data through the database was much faster and easier.

4.2.3 Data Mart

Concluded the ELT process, the final data mart was defined according to the Figure 4.9.

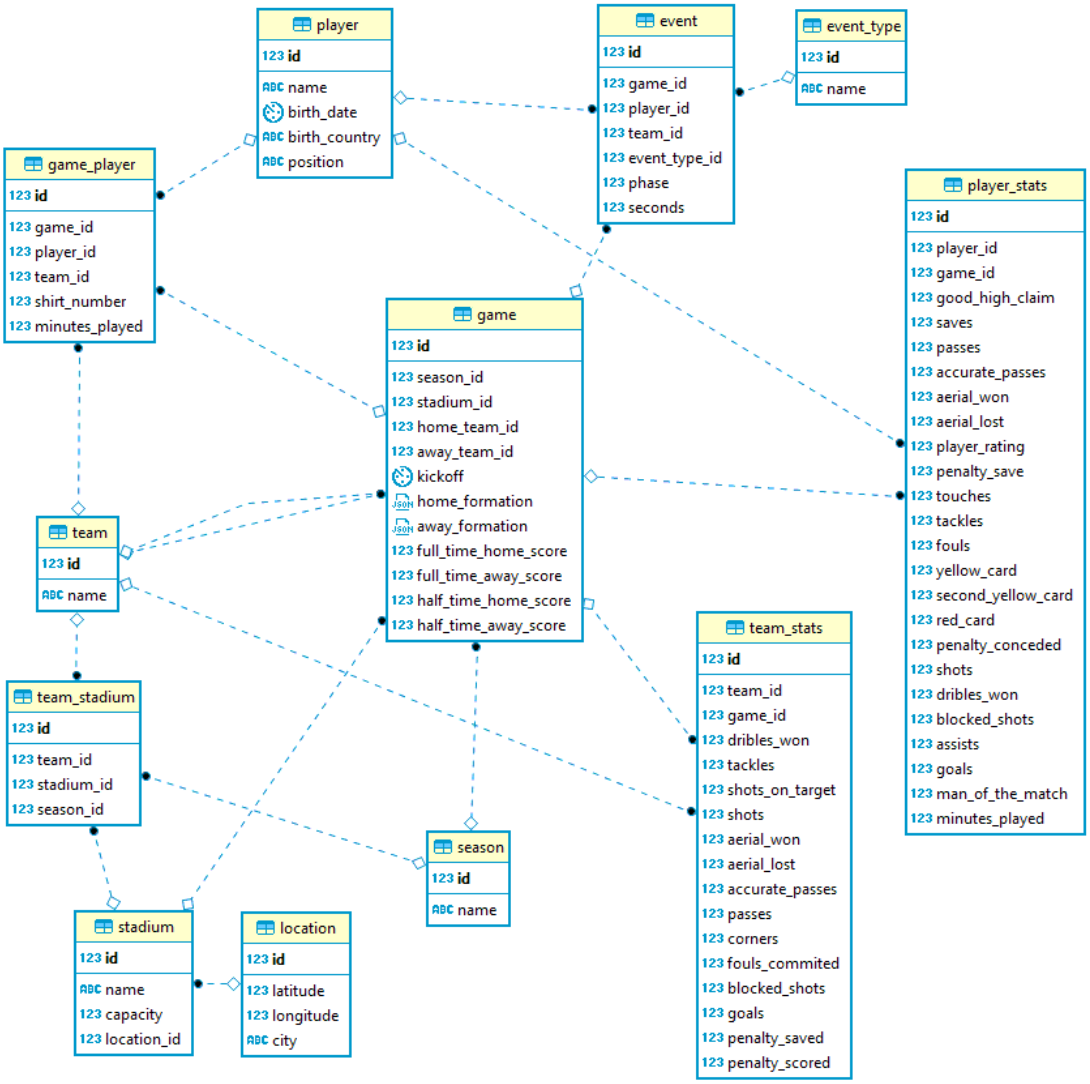


Figure 4.9: Entity relationship diagram of soccer data mart.

Chapter 5

Data Modeling and Predictions

One of the main objectives of this work is to predict game results, both in soccer and in basketball. This chapter details the implementation of the mechanisms applied to achieve this goal. The mechanisms are divided into ML for basketball and soccer, and in simulation for basketball.

For each mechanism is explained the models' creation and their results, including a comparison of results with other works.

5.1 Machine Learning

During the project several features for the data set have been thought, analyzed and selected to simplify and optimize the supervised learning algorithms applied. These algorithms were: decision tree; support vector machine; random forest; extreme gradient boosting (XGBoost); extreme randomized trees (described in section 2.2.1). Some to solve regression problems, others to solve classification problems. These details are explained throughout the chapter.

To apply the algorithms was used the Scikit-Learn package [89], which is a simple and efficient tool for data mining and data analysis in Python language.

Firstly, the basketball section is described, which was the first to be implemented. Lastly, soccer analysis is presented. Even though they are different sports, the problems encountered during the basketball analysis provided knowledge to the soccer one.

5.1.1 Basketball

In the analysis were used the seasons of 2012-2013 until 2014-2015 as the training data set and the season 2015-2016 as the test set. The games refer only to the NBA regular season, so the playoffs are not taken into account. As a result, the volume of games between seasons is regular, and it is possible to compare the analysis done with existing work. Each season has 1230 games, where each team plays 82 games. The only exception is in the 2012-2013 season since a bombing in the

Boston Marathon led to the cancellation of the game between Indiana Pacers and Boston Celtics. As both teams had the position of the playoffs decided the game was not rescheduled.

Feature engineering, feature selection, model evaluation, and comparison of results is described in the subsequent sections.

5.1.1.1 Feature Engineering

The own knowledge supported by prior research are behind the ideas for the features engineered. Each feature described is calculated for the home and away team. The feature vector consists of the difference between each feature's [1] team value that will play the game. One caution to take is not to use future data to forecast a game. As a result, a timeline is built so that each feature only uses the information prior to the game date. This timeline, known as forward chaining, prevents data leakage, which could happen when the data set has data points distributed chronologically. In these cases, nested cross-validation is performed using forward chaining method. Forward chaining involves creating many different folds in the data set so that the second half is predicted for each of these folds [90]. Figure 5.1 represents this process, where the blue points are the information from the games that happen before the game to be predicted (red point).

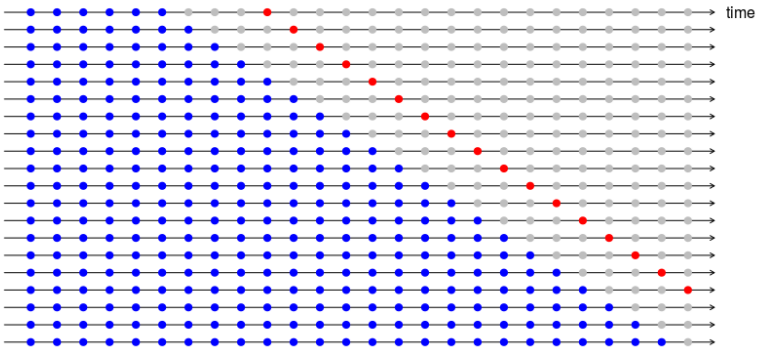


Figure 5.1: Forward chaining structure [90].

Another point is that the features are computed for each season independently, that is, the information does not go from one season to another. One limitation originated by the features' temporal dimension is the impossibility to forecast the first games due to lack of information. But to ensure a more accurate model, it is required to resort to this limitation.

Below is explained each feature, in some cases along with their mathematical formulations.

Basic Statistics

To keep track of what the team does more or less during games, basic statistics are used to describe the team offense and defense. The statistics are the following:

[1] Made and attempted field goals; made and attempted three point field goals; made and attempted free throws; offensive and defensive rebounds; assists; steals; blocks; turnovers; personal fouls; players weight; weighted streak; weighted streak location; rest days; form.

- Made and attempted field goals;
- Made and attempted three point field goals;
- Made and attempted free throws;
- Offensive and defensive rebounds;
- Assists;
- Steals;
- Blocks;
- Turnovers;
- Personal Fouls;

To take account of performance trends were taken the average values over the previous k games. The feature for a team's jth match is determined by:

$$\mu_j^i = (\sum_{p=j-k}^{j-1} \mu_p^i) \ / \ k \tag{5.1}$$

Where μ^i belongs to the set of basic statistics.

Players Weight

There is a transfer window during the season where the players change teams and, consequently, the strength of teams changes. To consider changes across the seasons, the players weight feature accounts for a player's performance relative to the time played. Mathematically, the expression is defined as:

$$GS\ (MP/GT) \tag{5.2}$$

Where the GS means game score, MP it's the time played, and GT it's the game's time. Game score is a metric created by John Hollinger to give a measure of a player's productivity for a single game [91]. The final value of the feature is the sum of the weight of each player in the team.

Weighted Streak

The growth or decrease in the team's performance is a very important factor. A team can start the championship very poorly derived from several factors and from one moment to another turn it around and achieve several consecutive wins. In order to accomplish good game predictions, it is essential to take into account the peak form of a team. For this matter was engineered a feature called the weighted streak, based on [38]. The feature encapsulates the results of the team in recent games, where a victory has more weight than a defeat. A temporal dimension was also included by placing time-dependent weights in the results of the previous games. In this way, a victory in a more recent game carries more weight than a victory in a game made some time ago. Since j

represents the team's game number and k the number of previous games, the weights are defined as follows:

- A weight of 1 on the oldest game ($j-k$);
- A weight of k on the most recent game ($j-1$).

Mathematically weighted streak for a team's jth match is defined as:

$$\frac{\sum_{k=1}^{j-1} kw}{\sum_{k=1}^{j-1} k} \tag{5.3}$$

Where w is one if the team wins the game, otherwise it is zero.

Weighted Streak Location

The comfort of players training regularly in their field along with the atmosphere provided by the fans offers the home team a certain advantage. As a result, the weighted streak location feature was created, which represents the team results in games at home or away. It is based on the weighted streak. The only thing that changes is the number of games because it counts only the games at home or away depending on the location of the game to predict.

Rest Days

The NBA championship is one of the most exhausting, where sometimes a team plays twice in two days. Therefore, the rest of the players between games is important because their performance decreases a lot when they feel tired.

In this way, one of the features corresponds to the number of days between games. For the first games, each team starts with a value of 4. This value is estimated from the beginning history of the regular NBA, which usually begins four days after the pre-season ends.

Form

To describe more strongly the trends of the championship the feature form was developed. It comprises a team's productivity relative to their opponents, unlike the weighted streak features. Based on the work [38], the feature ensures that a theoretically weaker team gets more credit when defeats a stronger team. For example, a win against a team on the same level is different than against a team with a higher level, and vice-versa.

At the start of the championship, each team begins with the form value at one and after each game, it is updated accordingly to the game result.

Throughout the competition they steal form from each other, that is, when team α wins against team β, team α takes a fraction of the form of team β. The form value (ξ) for each team's jth match is updated accordingly to the expression stated in [38]:

$$\begin{aligned} \xi_j^{\alpha} &= \xi_{(j-1)}^{\alpha} + \gamma\, \xi_{(j-1)}^{\beta} \\ \xi_j^{\beta} &= \xi_{(j-1)}^{\beta} - \gamma\, \xi_{(j-1)}^{\alpha} \end{aligned} \tag{5.4}$$

Where γ represents the stealing fraction and is comprised between $[0,1]$. When teams are on the same level (same form) the upgrade does not have a big impact, however when the levels are unbalanced the upgrade is higher. The stealing fraction was defined after cross-validating with different values of γ. Table 5.1 shows the optimized stealing fraction value for each model.

Table 5.1: Optimized stealing fraction value for each basketball model.

Algorithm	**Optimized Stealing Fraction**
Decision Tree	0.02
SVM	0.43
Random Forest	0.58
XGBoost Classifier	0.11
XGBoost Regressor	0.05
Extremely Randomized Trees	0.77

In short, Table 5.2 represents all features engineered among with an abbreviation to help a better visualization of the graphics' captions.

5.1.1.2 Feature Selection and Transformation

After defining which features to apply in the models, it is essential to pre-process the data. This pre-processing involves identifying the correlated features, interpret which algorithms behave better when the inputs are normalized or standardized, select features with more importance, among other details.

As a first step, was chosen to identify the features with the highest correlation, which means that they evolve in a similar way and may not be adding relevant information to the data set. Figure 5.2 represents the correlation between the variables, where the darkest cells correspond to the highest values.

Table 5.2: Features engineered for basketball.

Name	Abbreviation
Made Field Goals	MFG
Attempted Field Goals	AttFG
Made Three Point Field Goals	M3PFG
Attempted Three Point Field Goals	Att3PFG
Made Free Throws	MFT
Attempted Free Throws	AttFT
Offensive Rebounds	OR
Defensive Rebounds	DR
Assists	A
Steals	S
Blocks	B
Turnovers	T
Personal Fouls	PF
Players Weight	PW
Weighted Streak	WS
Weighted Streak Location	WSL
Rest Days	RD
Form	F

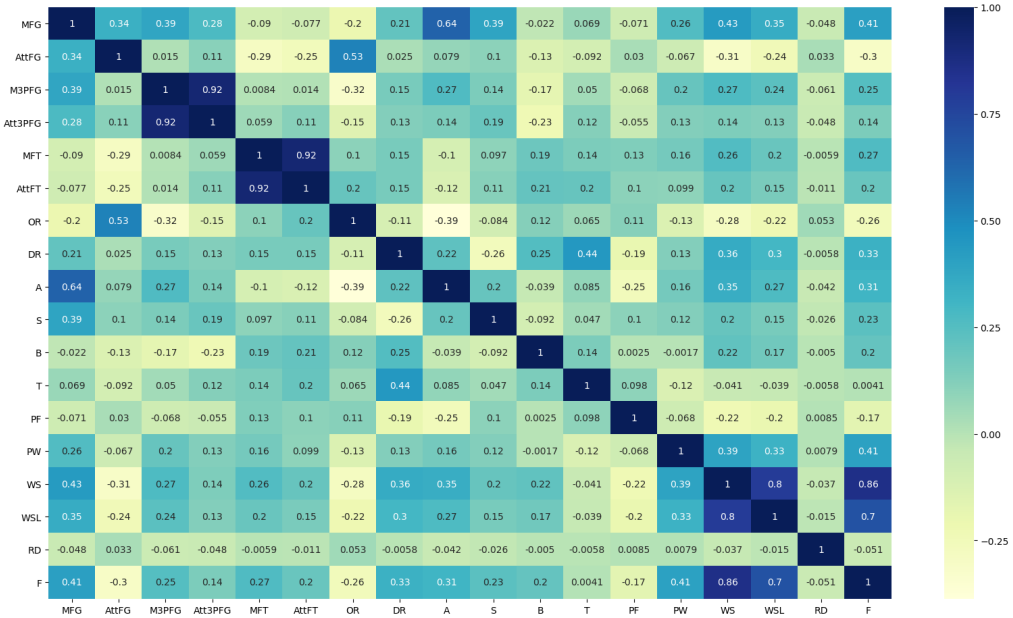


Figure 5.2: Features correlation from basketball data set.

In this filtering process, several limitation values of correlation were adopted, for example, a value of 0.9 or 0.8, in order to check if the performance of the algorithms increased. The feature selection resulted in better performance only in SVM, and the features removed were: attempted three-point field goals and attempted free throws.

On the other hand, some algorithms need to receive the inputs in a normalized or standardized form. Normalization re-scales the values of the data set to be between zero and one, which makes the training less sensitive to the scale of features. This method is also known as min-max normalization, and is implemented through the following equation:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{5.5}$$

The standardization re-scales the values so that the mean is equal to zero and the standard deviation to one, allowing to compare features that have different scales or units. Also know as standard score is calculated by:

$$z = \frac{x - \mu}{\sigma} \tag{5.6}$$

This data transformation was only applied in two algorithms. In the case of extreme gradient boosting regressor was applied a normalization, because the booster was not a tree, otherwise the algorithm would not need it. The other algorithm was SVM, because the way it works requires the data to be standardized in order to have a better performance [92]. Decision tree, random forest, and extremely randomized trees are tree-based models and hence do not require feature scaling.

5.1.1.3 Results

To forecast an entire season, the problem was solved using two methodologies: predicting the score difference between two teams and predicting the outcome of the game. The score difference is a regression problem, and for this methodology was applied the algorithm XGBoost regressor. The prediction of the outcome is a classification problem, in which the classes are win and lose, and for this problem the algorithms applied were XGBoost classifier, decision tree, random forest, extremely randomized trees, and SVM.

In addition to the feature selection, the hyper-parameter tuning was performed to optimize each model (5.3). To evaluate each combination of hyper-parameter values, it's necessary to optimize using a validation set. An important concept is not to tune the hyper-parameters on the testing data, because the purpose of the test set is to be used once, when the model is in the final state. For this purpose, using the Scikit-Learn package, a grid search was applied including the season 2014-2015 to validate. It was decided not to use k-fold cross-validation because it meant defining folds manually due to the timeline between games. Using an entire season as validation was the simplest process.

For the support vector machine, two kernels were used, namely the linear kernel and the radial basis kernel (RBF). During the optimization of the RBF kernel, the main hyper-parameters chosen were C and gamma. C represents the penalty parameter of the error term, trading off misclassification on training data against simplicity of the decision surface, and it is common to all

Table 5.3: Optimized hyper-parameters for each basketball model.

Algorithm	Optimized Hyper-Parameters
Decision Tree	max_depth = 3; max_features = 0.11; min_samples_leaf = 3; min_samples_split = 3
SVM	kernel = 'linear'; C = 50
Random Forest	max_depth = 8; max_features = 0.17; min_samples_leaf = 4; min_samples_split = 2; n_estimators = 100
XGBoost Classifier	booster = 'gbtree'; learning_rate = 0.01; reg_alpha = 0.05; reg_lambda = 0.025; n_estimators = 500; max_depth = 5
XGBoost Regressor	learning_rate = 1; n_estimators = 500; reg_alpha = 0.025; reg_lambda = 0.025
Extremely Randomized Trees	max_depth = 8; max_features = 0.23; min_samples_leaf = 4; min_samples_split = 4; n_estimators = 100

SVM kernels. Gamma is a kernel coefficient for non-linear hyperplanes [2] and defines how much influence a single training data point has [37].

The linear kernel produced better optimizations, for which the optimal value for the C parameter was 50.

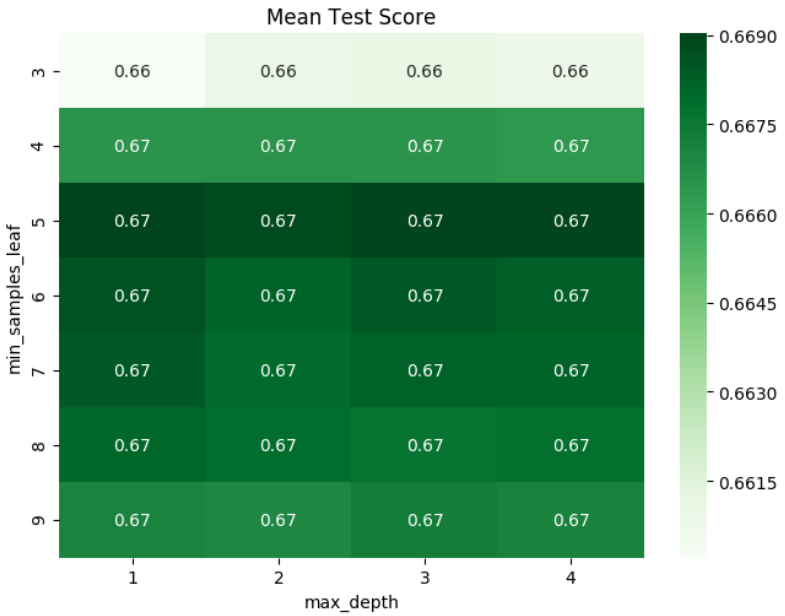


Figure 5.3: Tuning for min samples leaf and max depth in random forest applied to basketball.

The main hyper-parameters chosen to optimize in the random forest case were: max depth, max features, min samples leaf, min samples split, and estimators. Max features parameter corresponds to the number of features that the model takes into account when searching for the best split. Min samples leaf parameter is the minimum number of samples needed to be at a leaf node,

[2] subspace whose dimension is one less than that of its ambient space.

which could help to smooth the model, principally in regression. Minimum samples split parameter represents the minimum number of samples required to split an internal node. Lastly, the estimators' number parameter means the number of trees used in the model [37].

The optimized hyper-parameters correspond to the use of greater depth and a maximum of 17% of the feature set in each split. Figure 5.3 displays the results of the grid-search for the minimum samples at a leaf node and the maximum depth of each tree.

Since decision tree and extremely randomized trees are tree-based models, the hyper-parameters for the grid-search were almost identical, except for decision tree, which is a model based on only one tree hence it does not have the number estimators parameter. In terms of samples to be at a leaf node and samples to split an internal node, the optimal values for both algorithms were close, but the maximum depth in decision tree was lower than in extremely randomized trees.

Apart from the algorithms explained, it was used the library XGBoost [93], instead of using the Scikit-Learn package, to implement gradient boosting in the data set. Extremely gradient boosting follows the basis of gradient boosting but uses a more regularized model formalization to control over-fitting, thus solves problems in a fast and accurate way.

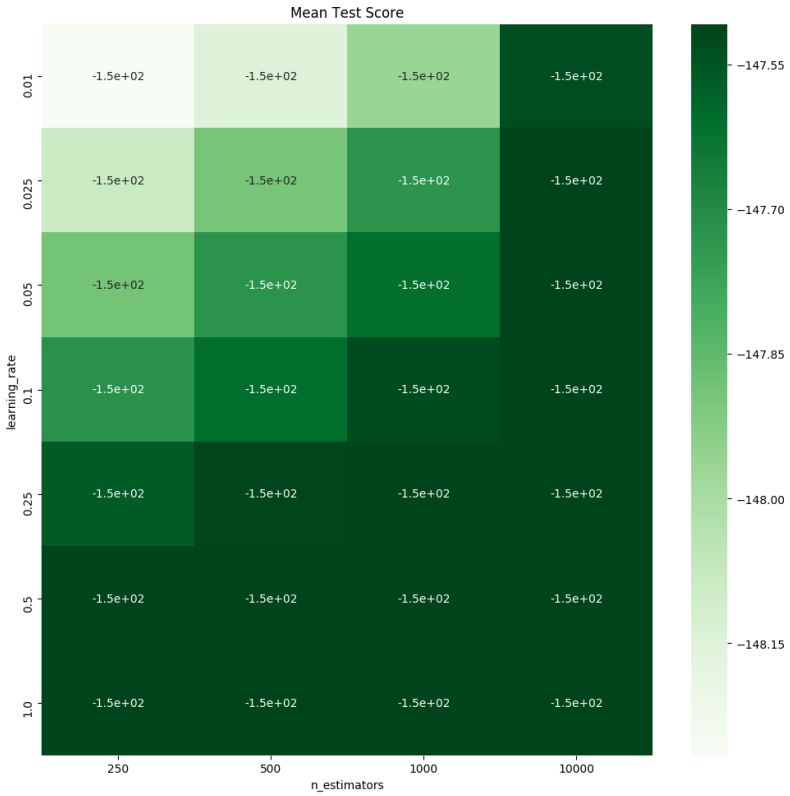


Figure 5.4: Tuning for learning rate and number estimators in XGBoost regressor applied to basketball.

To apply this machine learning technique were used two types of boosters, the gblinear and the gbtree, which use linear functions and tree-based models respectively. For both models were tunned the following hyper-parameters: learning rate, which shrinks the contribution of each tree,

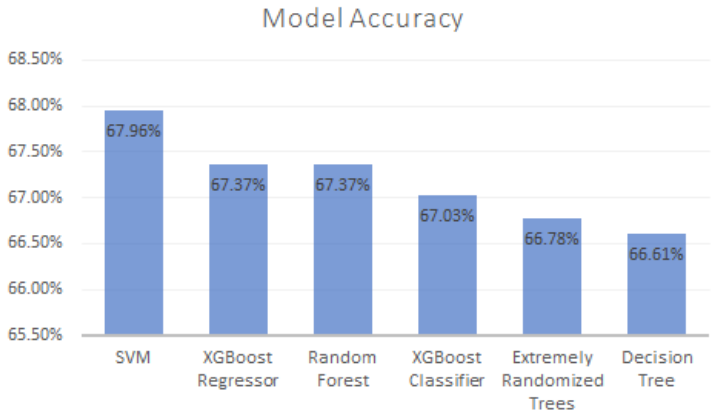


Figure 5.5: Accuracy scores of the different machine learning models applied in basketball.

number of estimators (linear or trees), alpha and lambda, that control the regularization term on weights [37]. In addition, the maximum depth of each tree was used in the tree booster. Figure 5.4 displays the results of the grid-search for the learning rate and the number of linear estimators.

Concluded the optimizations, the algorithms were applied to the test set and the score accuracy results obtained for each model is shown in Figure 5.5.

The best performing model was SVM, followed by random forest and XGBoost regressor, which showed the same accuracy, then XGBoost classifier, and extremely randomized trees. The worst performing model was the decision tree, but it is a model that allows better visualization of the chosen features that affected the predicted outcome.

Tables 5.4, 5.5, and 5.6 show that the three best performing models present high precision and recall values in the majority class (win class). Figure 5.6 illustrates the ranking of feature importance from random forest.

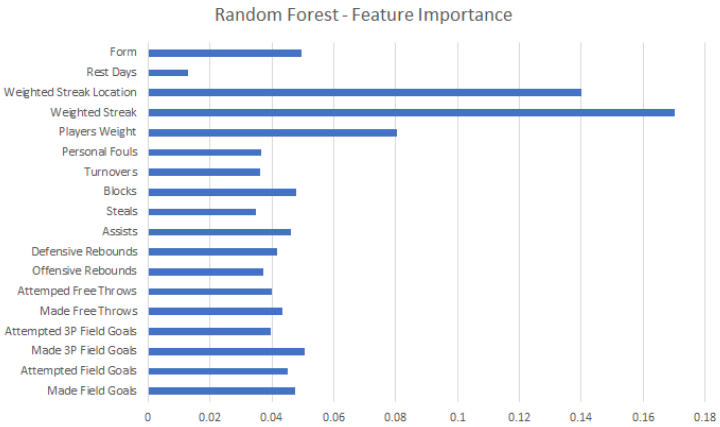


Figure 5.6: Feature importance from random forest applied in basketball.

Table 5.4: Basketball SVM results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Total
Home Win	556	150	706
Home Loss	231	252	483
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.71	0.79	0.75
Home Loss	0.63	0.52	0.57

Table 5.5: Basketball random forest results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Total
Home Win	558	148	706
Home Loss	240	243	483
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.70	0.79	0.74
Home Loss	0.62	0.50	0.56

Table 5.6: Basketball XGBoost regressor results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Total
Home Win	562	144	706
Home Loss	244	239	483
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.70	0.80	0.74
Home Loss	0.62	0.49	0.55

Table 5.7: Basketball extremely randomized trees results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Total
Home Win	601	105	706
Home Loss	290	193	483
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.67	0.85	0.75
Home Loss	0.65	0.40	0.49

Decision tree, despite being the model with the lowest accuracy, was the model that presented the best results in forecasting the loss of the home team, which was the hardest class to predict in all models, with precision and recall scores of 0.58 and 0.63 respectively.

Table 5.8: Basketball decision tree results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Total
Home Win	487	219	706
Home Loss	178	305	483
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.73	0.69	0.71
Home Loss	0.58	0.63	0.61

Table 5.9: Basketball XGBoost classifier results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Total
Home Win	553	153	706
Home Loss	239	244	483
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.70	0.78	0.74
Home Loss	0.61	0.51	0.55

Extremely randomized trees was the one that had the most variance between the two classes, and as a result, was the model that best predicted the home wins with a recall value of 0.85 (Table 5.7).

Finally, the XGBoost classifier had a variance between the two classes similar to the top performer algorithms but had a smaller amount of correct predictions.

A comparison was done with previous published works, in order to determine the added value of the methods used in this dissertation. The focus is on the performance of the algorithms, not so much on the data sets (seasons) used. Although, part of the analysis matches with work that predicts the same season, which gives a more direct comparison.

Table 5.10 displays the results achieve on Valenzuela's thesis [41]. The SVM and random forest models produced in this dissertation outperformed Valenzuela's results and even the other techniques used by him have achieved lower accuracy values than any of the models developed.

Table 5.10: Valenzuela's thesis results on NBA season 2014-2015 [41].

Algorithm	**Accuracy (%)**
Logistics	63.03
Naïve Bayes	61.76
Neural Networks	64.71
Random Forest	63.45
Stack	64.29
Adaboost	63.03

Table 5.11: Cao's dissertation results on NBA season 2010-2011 [40].

Algorithm	**Accuracy (%)**
Simple Logistics	67.82
Naïve Bayes	65.82
SVM	67.22
Neural Networks	66.67

Cao's dissertation predicts the season 2010-2011 presenting higher results than Valenzuela's work (Table 5.11). However, they are not enough to exceed the performance of all the models developed in this dissertation. Since the SVM exceeds all the results obtained by Cao.

The article [42] surpassed the best model developed, with 68.06% accuracy in the season 2008-2009 forecast using Naïve Bayes. However, the author used other models to predict several seasons, and in the case of the random forest, the best result was 64.91% (season 2010-2011), which did not exceed my random forest performance.

5.1.2 Soccer

For soccer, as the training set were used the seasons 2014-2015 until 2016-2017, and the season 2017-2018 as the test set. Each season has 380 games, where each team plays 38 games. Therefore, the volume of games between seasons is regular.

In the following sections is described the feature engineering, feature selection, model evaluation, and comparison of results.

5.1.2.1 Feature Engineering

The development of this stage was very similar to the features engineered in basketball (section 5.1.1.1) since some performance metrics and concepts of that analysis were successfully applied in soccer. Therefore, the feature vector describes the difference between the calculated features [3] values for the home and away team, and a timeline is built to prevent using post-game information during the prediction. The explanation and purpose of this timeline were explained earlier in the basketball's features.

In addition, the features are computed independently of the seasons, so there is no knowledge transmission from one season to another.

Below is presented each feature's concept, along with the mathematical formula.

Goal Difference

An important metric for predicting game outcomes is the goal difference. A method proposed by Constantinou and Fenton [94] reveals that this metric gives good results in soccer forecasts. That said, the goal difference was formulated, which is the difference between the sum of the goals scored and the goals conceded until the game.

For a team's jth match, using the k previous games, the goal difference is determined by:

$$GD_j = \sum_{k=1}^{j-1} GS - \sum_{k=1}^{j-1} GC \tag{5.7}$$

Where GS and GC represent, respectively, the goals scored and the goals conceded.

[3] Goals difference; corners; shots on target; goals; streak; weighted streak; weighted streak location; form.

Offensive Statistics

To represent the pressure applied on the opponent and the superiority of the team in terms of offense, the statistics of corners, shots on target, and goals were chosen. According to the paper [38], these are important statistics, because in the case of corners, if the number is high means that the team is forcing the opponent to dodge their offensive attempts. Shots on target demonstrate supremacy in reaching the opponent's soccer goal, and obviously, goals are fundamental since they determine the game result.

To take account of performance trends were taken the average values over the previous k games. For μ^i pertaining to the offensive statistics, the feature for a team's jth match is determined by:

$$\mu_j^i = (\sum_{p=j-k}^{j-1} \mu_p^i) \ / \ k \tag{5.8}$$

Streak

The streak feature covers the game result reached by the team in the previous k games. The streak value (δ) is computed through the team's league points secured in relation to the games made:

$$\delta_j = (\sum_{p=j-k}^{j-1} r_p) \ / \ 3k \tag{5.9}$$

Where $r \in \{0,1,3\}$, which represents the game result.

The feature is normalized so that the values are between zero and one. For this, the number 3 in the factor $3k$ of the formula 5.9 corresponds to the maximum of points that a team can get in a game. Thus, streak's maximum value is 1, meaning that the team has won all previous k games.

Weighted Streak

The weighted streak feature (described in section 5.1.1.1) only undergoes a small change in the formula 5.3. As in soccer, a game could end up tied, w belongs to the set $\{0,1,3\}$, which represent, respectively, lose, draw, and win outcome.

Weighted Streak Location

This feature applies the weighted streak feature concept described previously, but instead of using all the previous games, it filters them by location. Therefore, it shows different values when the game to predict is at home or away. The idea behind the feature is described in section 5.1.1.1.

Form

Section 5.1.1.1 has already described the concept and formulas of the form feature, but only when one team defeats the other. As in soccer there are draws it is necessary to formulate such

cases. In case of a draw, the feature developed by Babbota and Kaur [38], expresses the update of the form value as follows:

$$\begin{aligned} \xi_j^{\alpha} &= \xi_{(j-1)}^{\alpha} - \gamma\,(\xi_{(j-1)}^{\alpha} - \xi_{(j-1)}^{\beta}) \\ \xi_j^{\beta} &= \xi_{(j-1)}^{\beta} - \gamma\,(\xi_{(j-1)}^{\beta} - \xi_{(j-1)}^{\alpha}) \end{aligned} \tag{5.10}$$

Where γ represents the stealing fraction and is comprised between $[0,1]$. In this way, if a strong team (better form) draws with a weaker team (less form) loses form value, even though they did not lose the game.

Each model has an optimized value of stealing fraction, which was defined after cross-validating with different values. Table 5.12 shows the optimized stealing fraction value for each model.

Table 5.12: Optimized stealing fraction value for each soccer model.

Algorithm	**Optimized Stealing Fraction**
Decision Tree	0.33
SVM	0.01
Random Forest	0.84
XGBoost Classifier	0.36
XGBoost Regressor	0.93
Extremely Randomized Trees	0.2

In short, Table 5.13 represents all features engineered among with an abbreviation to help a better visualization of the graphics' captions.

5.1.2.2 Feature Selection and Transformation

The data pre-processing performed in basketball (described in section 5.1.1.2) was replicated for the soccer's work. Since the concepts and the reason to perform this step were described previously, this section focus on the changes coming from the soccer data set.

As can be seen in Figure 5.7, the correlation between features is lower, which is expected because the data set is composed of a smaller number of variables. Although there are two features with high correlation values, in this case, it was chosen not to filter the features by a maximum correlation value.

About feature scaling, since the algorithms applied in soccer are the same ones that were applied in basketball, that step was done in the same way.

5.1.2.3 Results

Same as basketball, the soccer season forecast followed an approach using the score difference of both teams as output (regression problem) and another approach that consisted in the classification of the game outcome (classification problem). Since soccer can have draws as an outcome, then there are three possible classes, thus moving to a multi-class classification problem.

Table 5.13: Features engineered for soccer.

Name	Abbreviation
Goals Difference	GD
Corners	C
Shots on Target	SoT
Goals	G
Streak	S
Weighted Streak	WS
Weighted Streak Location	WSL
Form	F

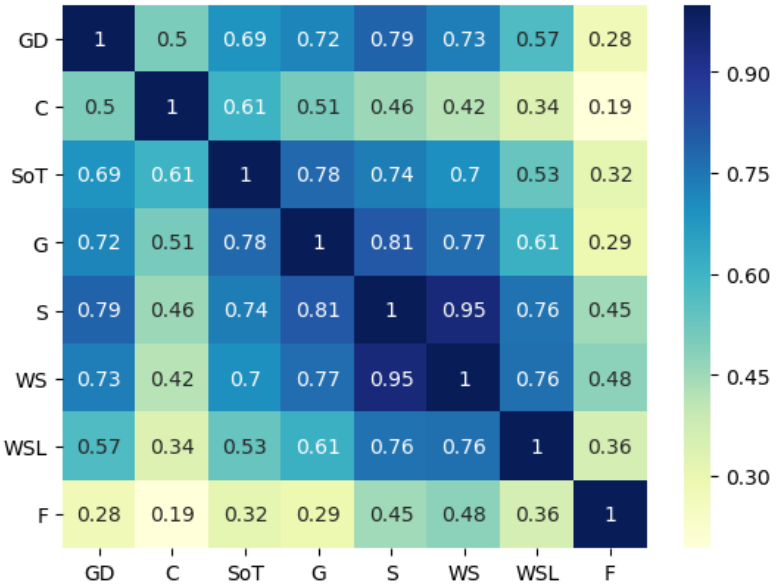


Figure 5.7: Features correlation from soccer data set.

In addition to the feature selection, the hyper-parameter tuning was performed to optimize each model. Chapter 5.1.1.3 shows the concepts behind grid-search, among with the hyper-parameters explanation for each model, so there is no need to explain them again. To optimize the hyper-parameters, using the Scikit-Learn package, a grid search was applied with the season 2016-2017 to validate.

In the case of random forest, it was found that the model behaves better when applying a smaller fraction of features in the best split. Consequently, maximum depth value of each tree and the minimum number of samples in each leaf node is also low. Figure 5.8 displays the results of the grid-search for the minimum samples at a leaf node and the maximum depth of each tree.

For the support vector machine, the best optimization was achieved with the RBF kernel using C and gamma values of 4 and 0.03 respectively.

The gradient boosting algorithms have the same optimized values, varying only on the alpha parameter. Figure 5.9 displays the results of the grid-search for the learning rate and the maximum depth of each tree in the XGBoost classifier.

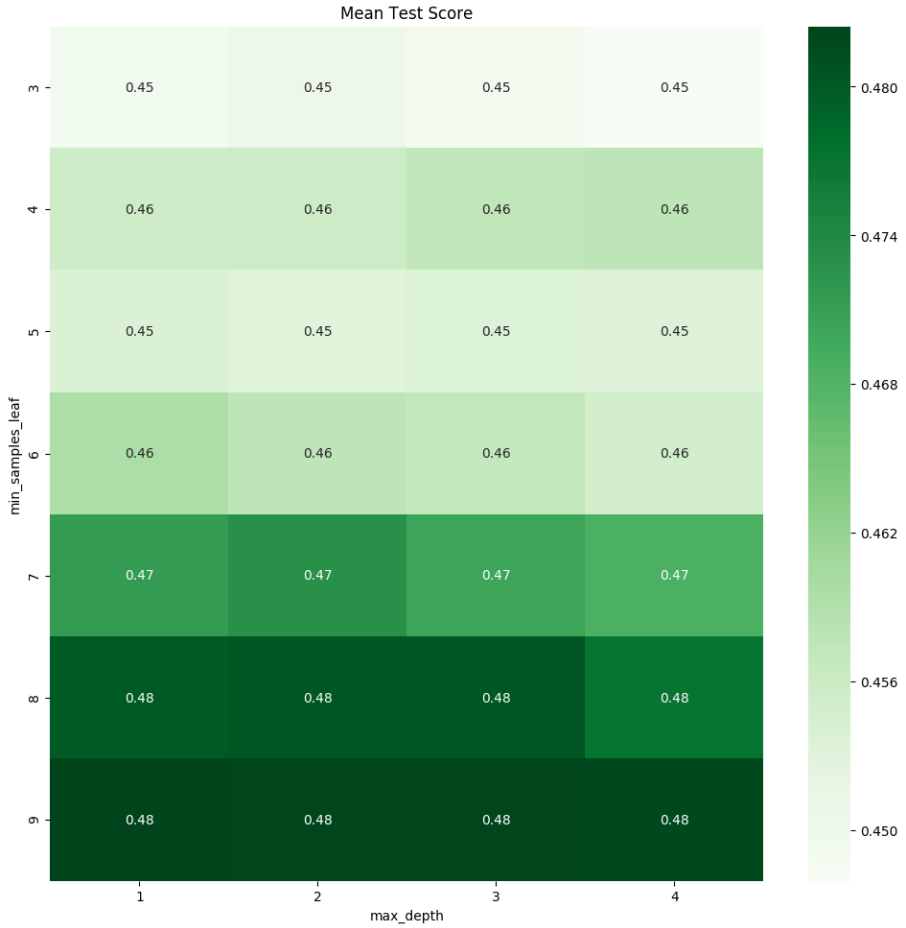


Figure 5.8: Tuning for min samples leaf and max depth in random forest applied to soccer.

Concerning the hyper-parameters of the tree-based models, the major difference between random forest and extremely randomized trees is the number of trees applied. In contrast, the decision tree has the values for samples at a leaf node and samples to split a node much higher than the other models. Table 5.14 describes all the optimized hyper-parameters for each model.

After tuning the hyper-parameters, the algorithms were applied to the test set and the score accuracy results obtained for each model is shown in Figure 5.10.

The best performing model was XGBoost regressor, followed by random forest, which was lower only by some tenths. Then, in order, decision tree, extremely randomized trees, and SVM. The worst performing model was the XGBoost classifier.

A large number of draws is one of the factors that make predictions in soccer so difficult, as this outcome is unlikely. Since 26% of games in the test set ended with a draw outcome, it is reasonable to achieve low accuracy results on the models. Another factor is the high competitiveness of the English Premier League and the wide occurrence of upsets. A great example was the season 2015-2016 (belongs to the training set), in which Leicester City won their first EPL title, against all possible predictions.

Although decision tree algorithm obtained one of the highest recall values for the win class, it

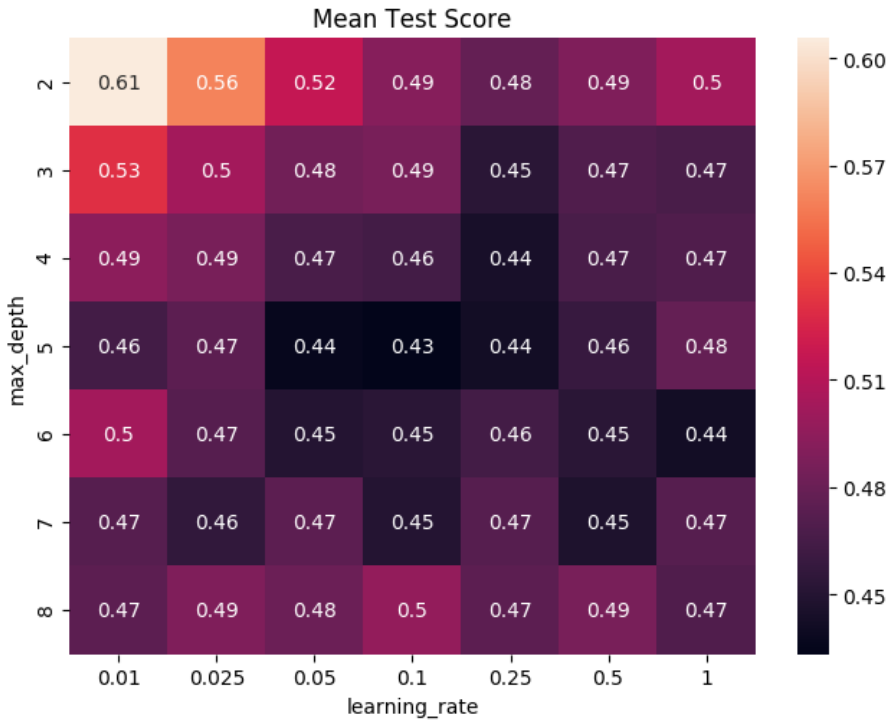


Figure 5.9: Tuning for learning rate and max depth in XGBoost classifier applied to soccer.

Table 5.14: Optimized hyper-parameters for each soccer model.

Algorithm	**Optimized Hyper-Parameters**
Decision Tree	max_depth = 5; max_features = 0.23; min_samples_leaf = 6; min_samples_split = 10
SVM	kernel = 'rbf'; C = 4; gamma = 0.03
Random Forest	max_depth = 4; max_features = 0.21; min_samples_leaf = 2; min_samples_split' = 3; n_estimators = 100
XGBoost Classifier	learning_rate = 0.01; n_estimators = 500; reg_alpha = 0.005; reg_lambda = 0.025
XGBoost Regressor	learning_rate = 0.01; n_estimators = 500; reg_alpha = 0.025; reg_lambda = 0.025
Extremely Randomized Trees	max_depth = 4; max_features = 0.25; min_samples_leaf = 4; min_samples_split' = 2; n_estimators = 1500

shows low precision, which means it had a high amount of true positives, but also false positives. The f1-score concerning the draw outcome reveals that the algorithm performed well in the toughest class. The f1-score related to the tie outcome reveals that the algorithm performed well in the toughest class to predict (Table 5.15).

The random forest was the worst algorithm at predicting draws with a f1-score value of 0.05, which is a metric that convey recall and precision in one. Consequently, it was the algorithm with higher values of recall for the others classes (Table 5.16). Figure 5.11 illustrates the ranking of feature importance from random forest.

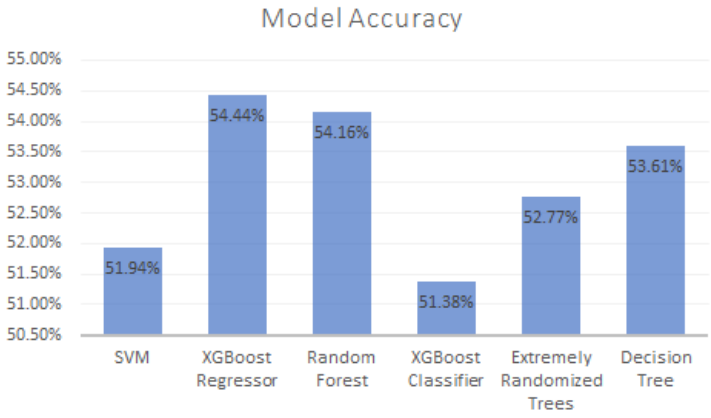


Figure 5.10: Accuracy scores of the different machine learning models applied in soccer.

As shown in Tables 5.17, 5.18, in terms of draws, gradient boosting models diverge greatly. The XGBoost classifier was the algorithm that predicted the draw class more often and obtained the highest recall value. Because of this more extensive attempt to forecast draws, it ended up failing more often in the other classes, which made it the model with lower performance. The XGBoost regressor has classified few draws, but those predicted were almost all correct, having a precision value of 0.89, the highest in that class. On top of that, it was the model with higher f1-score at predicting losses.

Both extremely randomized trees and SVM had similar results, having a medium precision/recall in predicting losses, and low precision and high recall for the win class.

The XGBoost regressor model needs the problem to be stated as a regression. The target variable used with this model was the score difference between the home team and the away team. In order to help determine a clear class when applying this model, the target variable was transformed into a standard score (Equation 5.6). Where x is the predicted score difference and the mean (μ) and standard deviation (σ) is applied to all values predicted.

After applied the transformation the values were classified in the following way:

- Win if the value is greater than 0.05;
- Tie if the value is between -0.05 and 0.05;
- Loss if the value is lower than -0.05.

A comparison was done with previous published works, in order to determine the added value of the methods used in this dissertation. Since the season used in the test set is from one year ago, it was difficult to find works that predicted this season. Thus, the comparison is based on the same models, but the predicted season will not be the same. It is not a limitation, because the focus is on the models performance, not so much on the data sets applied.

The article [44] predicted the seasons 2012-2013 and 2013-2014 of the EPL using several techniques. Two of them were SVM, applying an RBF kernel, and random forest, which achieved

Table 5.15: Soccer decision tree results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Predicted Draw
Home Win	126	23	15
Home Loss	45	49	6
Draw	62	16	18
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.54	0.77	0.63
Home Loss	0.56	0.49	0.52
Draw	0.46	0.19	0.27

Table 5.16: Soccer random forest results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Predicted Draw
Home Win	132	25	7
Home Loss	35	60	5
Draw	71	22	3
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.55	0.80	0.66
Home Loss	0.56	0.60	0.58
Draw	0.20	0.03	0.05

Table 5.17: Soccer XGBoost regressor results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Predicted Draw
Home Win	113	50	1
Home Loss	25	75	0
Draw	50	38	8
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.60	0.69	0.64
Home Loss	0.46	0.75	0.57
Draw	0.89	0.08	0.15

Table 5.18: Soccer XGBoost classifier results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Predicted Draw
Home Win	109	36	19
Home Loss	27	56	17
Draw	52	24	20
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.58	0.66	0.62
Home Loss	0.48	0.56	0.52
Draw	0.36	0.21	0.26

50.8% and 49.5% respectively. These results were outperformed by the models produced. Although Ulmer and Fernandez predicted more than one season, they also applied more 7 seasons than what were used to train the models of this dissertation, which ends up balancing the comparison. The best models developed also outperformed a neural network implemented by researchers from Slovak University of Technology [43], in which they achieved an accuracy result of 52.47%.

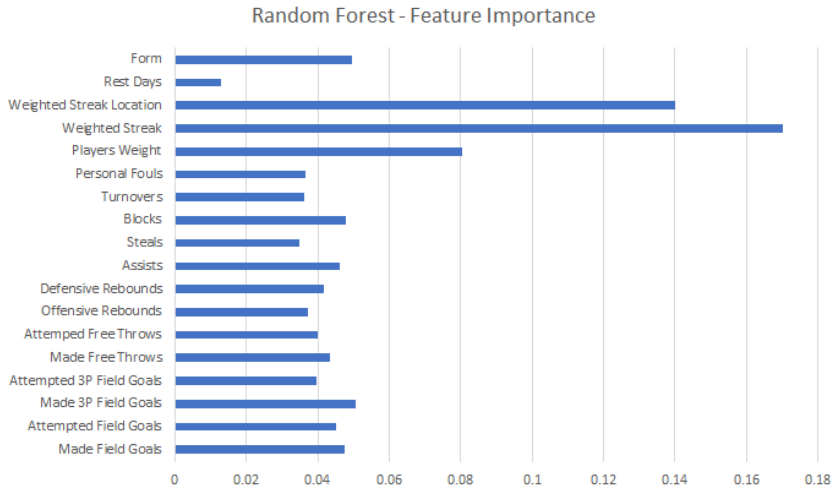


Figure 5.11: Feature importance from random forest applied in basketball.

Table 5.19: Soccer extremely randomized trees results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Predicted Draw
Home Win	124	24	16
Home Loss	33	55	12
Draw	60	25	11
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.57	0.76	0.65
Home Loss	0.53	0.55	0.54
Draw	0.28	0.11	0.16

Table 5.20: Soccer SVM results.

a) Confusion Matrix			
	Predicted Win	Predicted Loss	Predicted Draw
Home Win	117	29	18
Home Loss	37	53	10
Draw	63	20	13
b) Precision-recall			
	Precision	Recall	F1-score
Home Win	0.54	0.71	0.61
Home Loss	0.52	0.53	0.52
Draw	0.32	0.14	0.19

5.2 Simulation

A different approach to predict results is to use simulation. In this work, a simulation system for basketball was developed, in which the simulator model is based on probabilities extracted from the statistics of three NBA seasons. As for the ML models, the seasons of 2012-2013 until 2014-2015 were used for modeling and the simulations were performed in the fourth season (2015-2016), in order to predict the outcomes. The prediction of a game is determined by the team that has won the most simulated games. Figure 5.12 summarizes the simulation process.

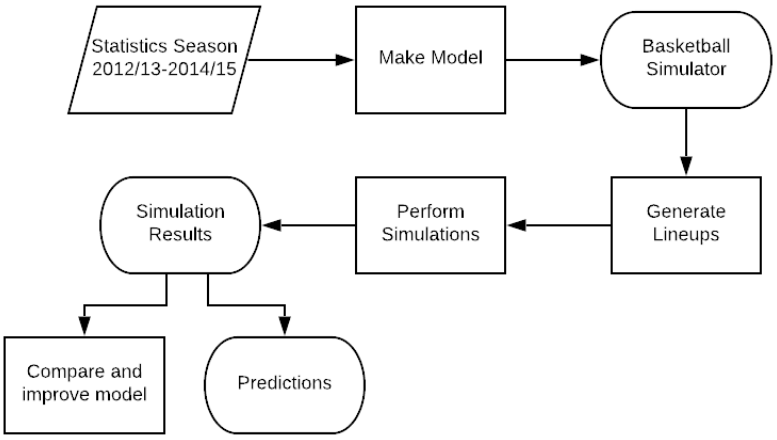


Figure 5.12: Overview of basketball simulation process.

One advantage of the simulation is the flexibility to change the lineups, in order to analyze what players perform better against a specific opponent.

During the section is described the features engineered, the implementation of the basketball simulator including the game actions flow, and finally, the analysis of the results achieved.

5.2.1 Feature Engineering

Each step of the simulation is performed using engineered features. Each feature is built, as earlier mentioned, from statistics from several seasons. This features model event probabilities that are used to select event outcomes or players. The features were all normalized according to equation 5.6 and were modeled as explained below:

Player Weight

The player weight feature is the same as the machine learning feature described in section 5.1.1.1. However, this feature is the value for each player and not the sum of the team. In this context, the feature helps generating the lineups.

Player Shot Frequency

The player likelihood (γ) to shoot is defined by the relationship between the attempted field goals ($PFGA$) from him and the attempted field goals ($TFGA$) from the team across the seasons, taking into account the played minutes made by the him (MP), as described in:

$$\gamma = (\sum_{k=1}^{j} PFGA_k * GT_k \ / \ TFGA_k * MP_k \ ) / \ j \tag{5.11}$$

Where GT represents the total game minutes. k and j represent the game number and the total number of games respectively. The minutes played were added according to the formulas of Basketball-Reference [91].

Choose Player To Shoot

To choose a player to attempt a shot, a probability for each player (p_i) is model. This probability is based on the article [58], and is defined according to the following equation:

$$\begin{aligned} p_i &= \theta_i + (\theta_i - \frac{1}{4} \sum_{k \in L_O, k \neq i} \theta_k) \\ \text{where } \theta_i &= \gamma_i - \sum_{j \in L_D} \beta_j \end{aligned} \tag{5.12}$$

The offensive lineup and defensive lineup are represented by L_o and L_d respectively. γ_i, which is calculated according to the equation 5.11, determines how likely a player is to take a shot and β represents the defensive ability of player j to prevent the player i from shooting. Also, it is important to model the shot frequency taking into account the teammates. To accomplish that is determined a team shot frequency coefficient $(\theta_i - \frac{1}{4} \sum_{k \in L_O, k \neq i} \theta_k)$ which can be negative if the average teammates' shot frequency is greater than the player i.

The defensive ability (β) includes the relation between the player defensive statistics and the defensive statistics of his team, and is model by:

$$\beta_{=}(\sum_{k=1}^{j} PS_k * PB_k * PDR_k \ / \ TS_k * TB_k * TDR_k \ ) / \ j \tag{5.13}$$

Where PS, PB, and PDR represent, respectively, the steals, blocks, and defensive rebounds made for a player during a game (k). The TS, TB, and TDR are the same type of statistics, but from the team.

Offensive Event

To model the probabilities of the offensive actions performed by a player, five events were identified:

- Score two points shot;
- Score three points shot;
- Fail two points shot;
- Fail three points shot;
- Lose the ball.

The probabilities are the average of these actions across the seasons. The events related to the shot take into account the shots attempted, while the event of losing the ball only takes into account the average turnovers of the player.

Before determining whether the shot was successful or unsuccessful, first the type of shot is selected, that is, whether it is two points or three points. The selection is formulated by the ratio of the player's attempted shots to the team's attempted ones.

One limitation of the model is that the statistics do not mention how close the defensive opponent was from the player. For example, a player who made 70% of the shots when the defender was around is more likely to score than a player who converted 80% of the shots, but without opposition from the defender. However, the model will favor the second player.

Rebound

To determine whether the player grabs an offensive or defensive rebound, the probability is model based on the idea of Oh et al. [58], in which the rebound is a competition between the players on the court. This way, given the offensive and defensive lineup on the court, the probability of player i catching a defensive or offensive rebound is defined by:

$$p(DR_i = 1|L_d, L_o) = \frac{\rho_i^d}{\sum_{j \in L_d} \rho_j^d + \sum_{k \in L_o} \rho_k^o} \tag{5.14}$$

$$p(OR_i = 1|L_d, L_o) = \frac{\rho_i^o}{\sum_{j \in L_d} \rho_j^d + \sum_{k \in L_o} \rho_k^o} \tag{5.15}$$

DR_i and OR_i represent the probability of player i catching a defensive and offensive rebound respectively. Since there is a difference between grabbing a defensive or offensive rebound, for each player i it was used the defensive rebound percentage (ρ_i^d) and offensive rebound percentage (ρ_i^o). Both metrics are the average of these values across the three seasons referred previously.

Number of Ball Possessions

To model the maximum number of ball possessions for a game is drawn random samples from the probability density function of the normal (Gaussian) distribution, which is defined by:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.16}$$

Where μ and σ are, respectively, the mean and standard deviation of the estimated number of ball possessions across the seasons. The estimated number of ball possessions is calculated based on the statistics from the two teams, and then averaged to provide a more accurate estimate [91].

5.2.2 Implementation

The simulator was implemented in Python language and behaves as a state machine. Every game action represents a state, which has specific pre-conditions according to the basketball game rules. Figure 5.13 demonstrates every action implemented and the flow of the state machine.

As said before, every state applies some feature engineered. Apart from the other states, the jump ball does not use training data and functions as a coin flip, because the data set does not provide jump ball statistics.

The simulation finishes when the maximum number of ball possessions is reached. If the game is tied, a new number of ball possessions is added to represent the overtime. The number of overtime ball possessions (OBP) takes into account the real time of a basketball game and is calculated by:

$$OBP = (BP * 5)/48 \tag{5.17}$$

Where BP represents the game ball possessions calculated when the simulation began. The 5 corresponds to the additional five minutes gave in an NBA overtime, and 48 corresponds to the length of an NBA game.

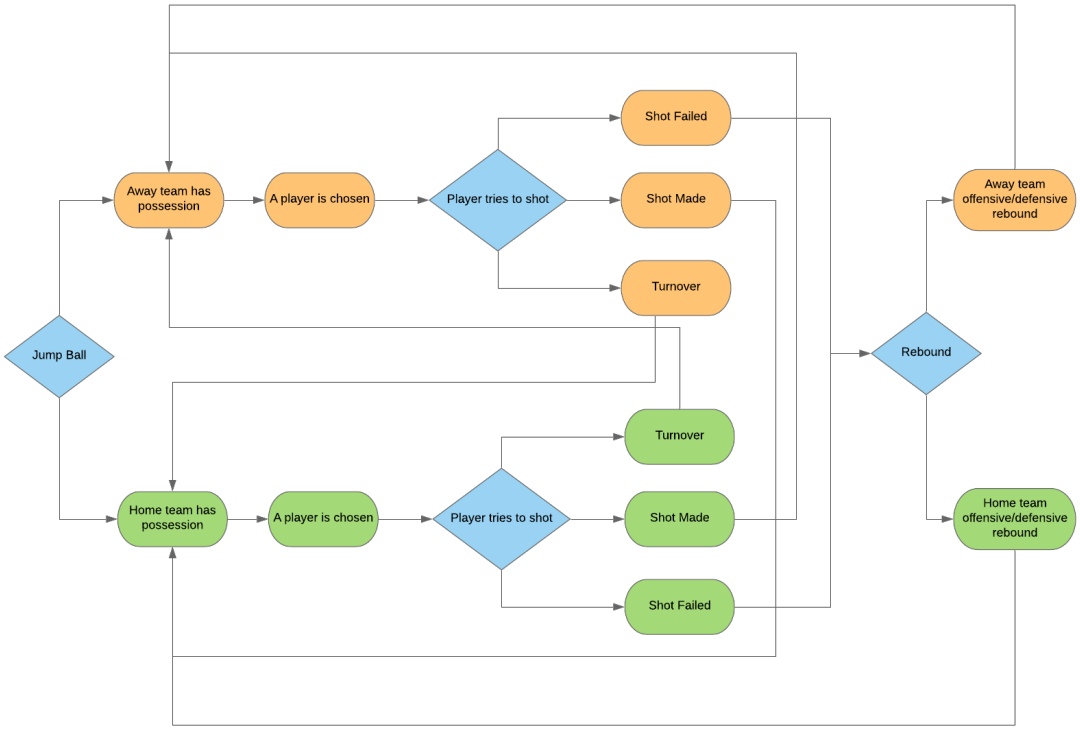


Figure 5.13: Basketball simulation flow.

For each action performed the simulation system saves who performed the action and then uses it to compute the statistics box score in the final of each simulated game. Most of the model improvements were identified through these box scores.

5.2.3 Results

The analysis of the model results was performed in two distinct phases. First, predict the game outcome through multiple simulations using a specific lineup as input. Second, the simulation of

an entire season, in which the prediction of each game is defined through several simulations with randomly generated lineups.

In basketball, it is common an athlete playing more than one position because some positions share the same characteristics. It was decided to make pairs of positions during the process of generating the lineups. Thus, it was fulfilled the gap of some teams having few players in certain positions, and the lineups became more diversified.

For each slot of the lineup, the following pairs were used: point guard (PG) and shooting guard (SG); PG and SG; SG and small forward (SF); power forward (PF) and center (C); PF and C. From a vector with the weights of the players according to the positions, it was randomly chosen a player for each slot.

For the first phase, the game between Cleveland Cavaliers and Chicago Bulls was simulated a thousand times, and the results were the following:

Table 5.21: Average statistics box score of an NBA game simulated a thousand times.

Cleveland Cavaliers											
	FG	FGA	FG%	3P	3PA	3P%	DRB	ORB	TRB	TO	PTS
LeBron James	14.1	26.8	.53	2.1	5.9	.36	10.6	2.1	12.6	5.0	30.3
Kevin Love	9.6	22.6	.43	3.2	8.9	.36	15.2	4.2	19.3	3.0	22.5
Timofey Mozgov	6.8	12.6	.54	0.03	0.3	.13	10.6	5.8	16.3	2.8	13.6
J.R. Smith	8.9	21.6	.41	4.0	10.9	.37	6.5	1.0	7.5	2.3	21.9
Mo Williams	8.0	19.5	.41	2.5	7.0	.36	4.1	0.9	5.0	4.2	18.5
Chicago Bulls											
	FG	FGA	FG%	3P	3PA	3P%	DRB	ORB	TRB	TO	PTS
Jimmy Butler	7.8	17.7	.44	1.4	4.4	.33	5.9	2.6	8.5	2.2	17.1
Pau Gasol	11.0	22.8	.48	0.3	0.9	.36	14.0	4.2	18.3	3.6	22.3
Nikola Mirotic	7.7	19.6	.39	3.6	10.4	.34	10.6	2.4	12.9	3.0	19.0
Derrick Rose	11.2	27.5	.41	1.9	6.4	.30	4.6	1.3	6.0	4.9	24.4
Tony Snell	5.3	13.8	.39	2.4	7.0	.34	6.4	0.9	7.2	1.9	13.1

The outcome predicted the victory of Cleveland Cavaliers, since they achieved an average score of 106.7 points, while Chicago Bulls scored an average of 95.8 points. Cavaliers gathered a total of 743 wins, and Chicago only had 257 during the 1000 simulations.

Table 5.22: Game results between Chicago Bulls and Cleveland Cavaliers during the season 2015-2016.

Date	Chicago Bulls	Cleveland Cavaliers
28-10-2015	**97**	95
24-01-2016	**96**	83
19-02-2016	95	**106**
04-10-2016	**105**	102

Comparing the simulation with the reality, the player level is well represented, as players who have scored the most points, such as LeBron James and Derick Rose, are players that between 2012 and 2015 have been selected at least once for the roster of the NBA All-Star. At the team level, as we can see in the table 5.22, the balance between the two teams is well represented (average difference of 10 points in the simulations), but the number of wins fails by a large margin. Chicago Bulls won 3 of 4 games (75%), while in the simulations they had a win rate of 27.5%. This difference is understood, since lineups are fixed and there are no substitutions, which means that the bench players from a team do not contribute to the predicted result.

For the second phase, several simulations for each game were performed to predict an entire season. The process consisted of generating five different lineups for each game and for each lineup simulate the game one hundred times. To select the players for the home and away team was used the average player's weight (described in section 5.2.1), therefore a player with a higher weight has higher probabilities of being selected.

The model achieved an accuracy of 64.80%, which is a satisfactory result considering some limitations of the system. Tables 5.23 and 5.24 respectively demonstrate the confusion matrix of the model and the wins that each team finished the simulation.

Table 5.23: Confusion matrix of the simulation model.

	Predicted Win	Predicted Loss	Total
Real Wins	477	247	724
Real Loss	186	324	506

Comparing the simulator predictions with ESPN Forecast panel [95] predictions, some teams do not differ much from the values of the simulation, such as Dallas Mavericks and Denver Nuggets, but others differ greatly, like New Orleans Pelicans.

Compared with reality, a satisfying fact is that the model hit all the teams that went to the playoffs.

Overall, strong teams like the Warriors and Cavaliers are extremely favored, which means that few upsets happen in the simulations.

Table 5.24: Simulation results of the NBA regular season 2015-2016.

Team	Wins
Golden State Warriors	79
San Antonio Spurs	77
Cleveland Cavaliers	76
Toronto Raptors	72
Oklahoma City Thunder	68
Los Angeles Clippers	63
Atlanta Hawks	63
Miami Heat	56
Charlotte Hornets	55
Boston Celtics	53
Indiana Pacers	48
Detroit Pistons	46
Portland Trail Blazers	45
Memphis Grizzlies	42
Dallas Mavericks	40
Chicago Bulls	37
Washington Wizards	35
Houston Rockets	34
Utah Jazz	33
Orlando Magic	33
Denver Nuggets	29
Milwaukee Bucks	25
Sacramento Kings	24
New York Knicks	24
New Orleans Pelicans	20
Minnesota Timberwolves	19
Phoenix Suns	14
Brooklyn Nets	13
Los Angeles Lakers	5
Philadelphia 76ers	2

Table 5.25: ESPN Forest panel predictions of the 2015-2016 NBA regular season [95].

Team	Wins
Golden State Warriors	60
Cleveland Cavaliers	56
Houston Rockets	56
San Antonio Spurs	55
Oklahoma City Thunder	55
Los Angeles Clippers	55
Atlanta Hawks	51
Chicago Bulls	50
Memphis Grizzlies	49
Miami Heat	46
Washington Wizards	46
New Orleans Pelicans	46
Toronto Raptors	45
Boston Celtics	43
Milwaukee Bucks	43
Utah Jazz	41
Indiana Pacers	39
Dallas Mavericks	38
Detroit Pistons	35
Charlotte Hornets	34
Phoenix Suns	36
Sacramento Kings	34
Portland Trail Blazers	31
Orlando Magic	30
Brooklyn Nets	28
Denver Nuggets	27
Los Angeles Lakers	27
New York Knicks	27
Minnesota Timberwolves	25
Philadelphia 76ers	19

Chapter 6

Data Visualization

One of the goals of the project is to provide a way to visualize the data extracted to support decision-making processes. This data visualization is divided into two different target audience, the team's staff, through reports with information from the opponent, and the spectators, where the information is presented through overlays in the game transmission.

This chapter explains the implementation of reports and overlays, including how the data that feed them has been organized and finally demonstrates a proof of concept of both types of visualization.

6.1 Feature Engineering

The purpose of reports and overlays is to filter an extensive amount of data into useful information and pass it effectively to the target audience. Due to that, were created database views that filter and organize information. Below is explained in detail each view.

Last 10 games of a team

Useful information about the performance of a team in the last days are the results obtained during that period. This data set contains the outcome, location, date, and opponent of the last ten games. The structure of the view is the same for basketball and soccer.

Score per game time

The score per game time is a good metric to analyze in which period of the game a team tends to score more. Due to physical wear and tear, or a strong physical preparation, there are teams that are more productive at the beginning or end of the game.

In the case of soccer, the view is composed by the average goal per game minute of a team. For example, in the season 2014-2015, Arsenal averaged 4 goals in the 63^{rd} minute.

In the case of basketball, the database view consists of the average points scored in a single quarter by a team.

Team Statistics

The team statistics represent the sum of offensive and defensive values, for example, assists, steals, shots, rebounds, among others.

For this case, two views are generated, one with the statistics until the moment of the game and another with the statistics of an entire season.

Player Highlight

Player highlight consists of presenting the player to watch for each position. The selection criterion is the average values of the player rating variable, in case of soccer, and the average values of the game score variable, in case of basketball. Depending on the position of the player, the statistics relevant to that position are shown. For example, for a goalkeeper, it is computed the rating and also the sum of his saves.

Equal to the team statistics, two views are generated, one with the player highlight until the game and another for the whole season.

6.2 Reporting

6.2.1 Existing Tools

For the generation of the reports, first, a research was done on the existing tools in the market. Since there is enough diversity, this section will only focus on three tools: Microsoft Power BI [96], Tableau [97] and QlikView [98].

Power BI

Power BI is a business analytics service designed for business users to access and combine different data sets for visualization and interactive dashboards. This tool comes with built-in data preparation and multiple features, such as natural language query and machine learning driven "Quick Insights" for automated discovery. In addition, it is possible to share with co-workers customized dashboards and interactive reports on any device.

Tableau

Tableau is known as a set of business user-friendly analysis and data visualization tools. Built-in intelligence and memory utilization to optimize performance are key components that make this tool so popular for self-service BI and data discovery cases. Due to the mapping functionality, it is possible to group data by geography, and also plot latitude and longitude coordinates.

Figure 6.1: Comparison between Tableau and PowerBI [99].

QlikView

QlikView is a tool for dashboarding and exploratory data analysis on in-memory technology and was one of the first products appearing on the market, related to the data discovery category. It is guided to help any business user to analyze data, without asking for a new report done by an expert.

Table 6.1 shows the results of a survey, concerning the use of each tool grouped by task. Although QlikView performs well for some tasks, the reporting category has a lower percentage of usage compared to the other tools. In this way, QlikView was discarded and a comparison between Power BI and Tableau was investigated further. The two tools are quite popular and demonstrate high performance in reporting, but based on Figure 6.1, Power BI was chosen to create reports for basketball and soccer.

Tasks/Product	Power BI	Tableau	QlikView	Average of all products
View	**91%**	89%	88%	91%
Navigate	85%	81%	**86%**	81%
Explore/Analyze	81%	76%	**82%**	79%
Create reports	46%	**51%**	33%	52%
Model/Enrich	**41%**	33%	21%	31%
Sample size	170	95	199	

Table 6.1: Tasks carried out with BI tools by business users [99].

6.2.2 Proof of Concept

As proof of concept, only soccer reports are exhibited. Starting with the team report (Fig. 6.2), the purpose is to show the results of the last 10 games of the season and the information about the team's goals. At the goal level, side-by-side are the goals scored and conceded, to give a thought about the team score difference. In addition, a chart with the average number of goals per game minute is also displayed.

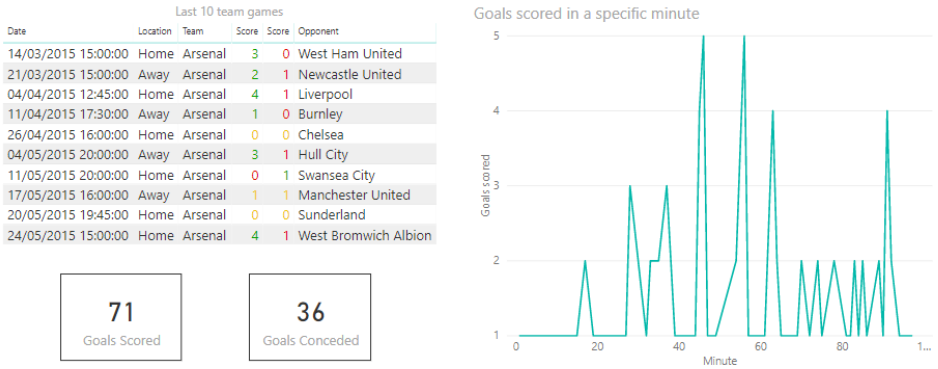


Figure 6.2: Soccer report with last games and goals from a team.

To get a closer look at team performance, the sum of several statistics is displayed, organized by category to make it easier to read, as can be seen in Figure 6.3.

Figure 6.3: Soccer report with team statistics.

Finally, a report that shows the most valuable players of a season, providing insight to coaches about the opposing players that he should observe more closely (Fig. 6.4).

6.3 Multimedia Augmentation

About the overlays, the purpose of the project is to provide the spectator with pre-game information that is relevant during the transmission of the game. An important aspect is that the overlay

Figure 6.4: Soccer report with most valuable players of a season.

needs to be displayed without there being a perceptible intrusion in the main content.

This chapter describes the system architecture that allows the insertion of overlays in the transmission and a proof of concept, which shows some examples of overlays and how the broadcaster controls what is displayed during the game.

6.3.1 Architecture

The implemented multimedia augmentation system consists of three components: RESTful API, server, and client.

The RESTful API, also described as RESTful web service, integrates sports data and provides various types of HTML5 overlays that include the data referenced previously. REST, which states for Representational State Transfer, is a software architectural style applied in web services communications. REST is more suitable for internet usage because it leverages less bandwidth [100].

The server, which triggers the client, aims to make the sports game live streaming. Live streaming consists of streaming digital content in real-time through the internet. The server uses the Flask framework [101], which is a micro web framework build for Python, and the Video-JS library [102] to support HLS. HLS, which means HTTP Live Streaming, is one of the most common transport protocol in a web context, and it was applied for the live streaming functionality.

The client controls which types of overlays want to display on the stream, and at what time.

One advantage of the system is that any client in any location can request the overlays, and the internet connection is the only requirement. Figure 6.5 presents the system architecture.

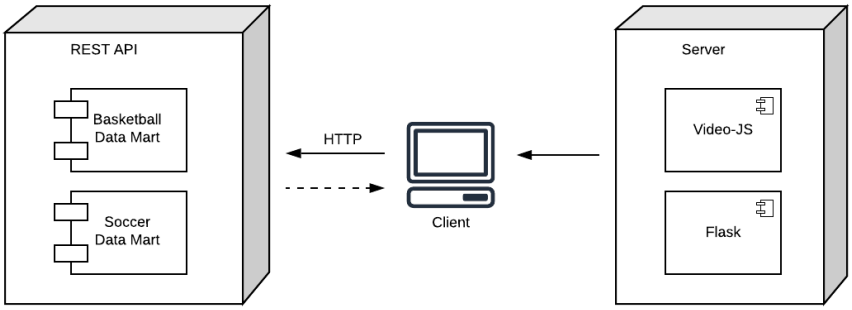


Figure 6.5: Multimedia augmentation system architecture.

6.3.2 Implementation

On the implementation, the server using the video-js framework only needs the manifest file (m3u8 extension) to transmit the game stream, which can be through the URL where it is hosted. The manifest file describes what segments of the live streaming are available, which audio languages the content has, and the different audio and video qualities available. The flask framework is used to develop a strong web application base.

Concerning the web service, each API's endpoint corresponds to one type of overlay, in this way any client, from any machine, can request a sports overlay. The overlay generation is dynamic considering that the web service only needs to receive as parameters the sport and the clubs that are going to play. Then, extracts the necessary information using those parameters, generates the overlay, and finally returns the HTML to the client. The overlays are implemented using HTML5 and CSS and are just an HTML layer that is placed above the HTML where is displayed the video.

The web application (client), at the beginning of the session, through javascript sends HTTP asynchronous requests to the RESTful API to get all the overlays available. This way, when the requests succeed, the client only needs to change the CSS dynamically in order to display the respective templates. The CSS changes occur through javascript code invoked by the user-interface (toolbox).

6.3.3 Proof of Concept

As proof of concept of the system, a demonstration is presented only with soccer data. The overlays developed are divided into two categories: team and player. For each category, there is a sub-category related to the image size. The small size allows presenting the overlay while the game is running, without there being an intrusion in the main content. On the other hand, the large size can only be displayed during game stops, because being larger can make it difficult to watch the main content.

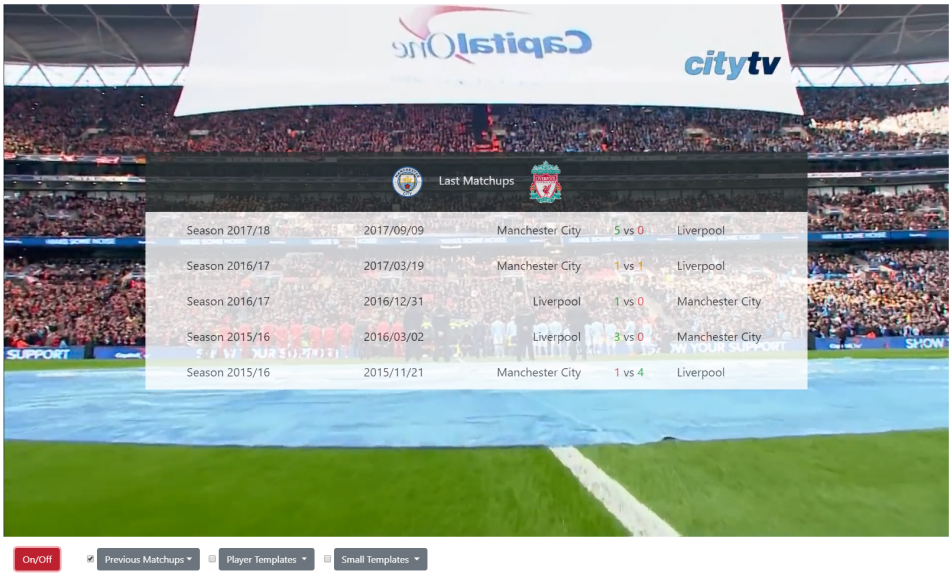


Figure 6.6: Soccer overlay with last matchups.

The team category includes overlays presenting the last matchups between the two teams (Fig. 6.6) and statistics for each team, in which could be the statistics from the entire previous season or from the current season until the current game (Fig. 6.7).

Figure 6.7: Soccer overlay including the team's statistics until the game.

The player category includes overlays presenting a comparison between the players performing best in each position, and the overall best player from each team.

The first one, display statistics concerning the position, for example, for forwards shows the goals and shots. As can be seen through Figure 6.8, it is a large overlay, so it must be used during game stops.

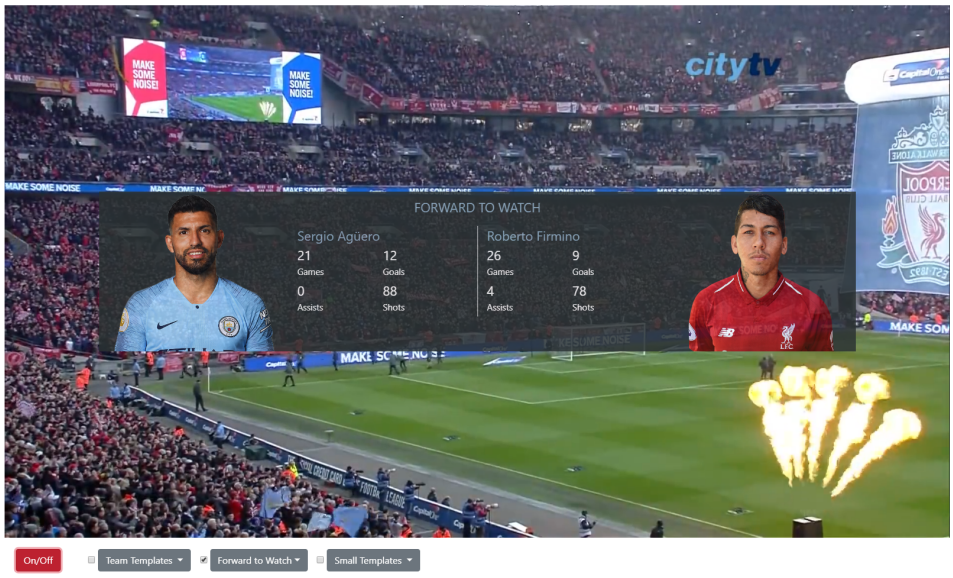


Figure 6.8: Soccer overlay including player comparison by position.

The second one only displays the player information, therefore it can be used when the game is running (Fig. 6.9).

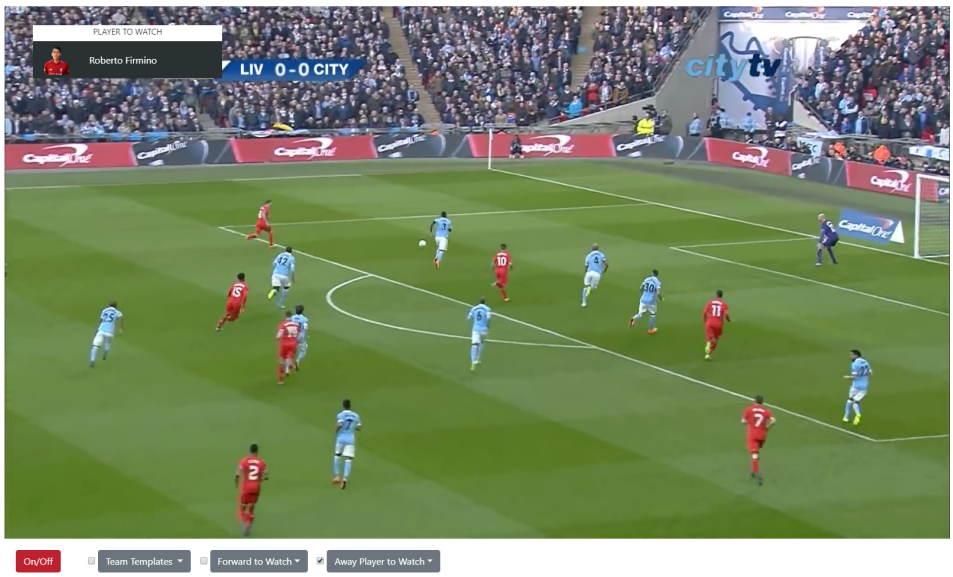


Figure 6.9: Soccer overlay including the overall best player of each team.

The user, through the toolbar below the video (Fig. 6.10), can manage which overlay's template to use and when to display. Drop-down lists allow to select a template to be ready, and it is possible to choose multiple overlays. The only constraint of multiple selections is that it can only be a template of each drop-down to prevent overlap. To trigger the selected template it is necessary to activate the check box on the left side. The red button enables to show or hide all overlays at once.

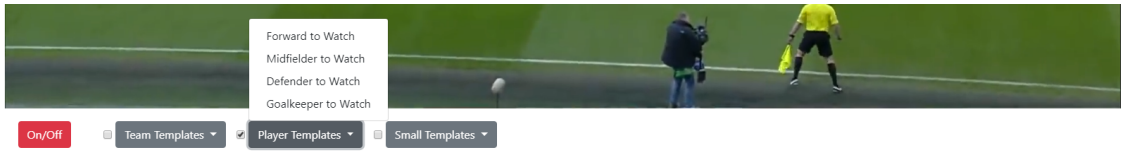


Figure 6.10: User interface to manage the overlays.

Chapter 7

Conclusions and Future Work

The existence of large amounts of information, and the distribution and massification of platforms for multimedia content visualization have given a great prospect for the consumption of these contents, namely inside a sports related environment. Sports analytics tools, that help teams winning championships or to enhance viewer experience, constitute a growing market that is becoming extremely popular.

Throughout this dissertation it was implemented a full business intelligence pipeline in order to extract more value from the collected data. After this integration of data into a multidimensional database, several layers were applied to the data: Business Intelligence Reporting, Machine Learning, and Simulation. The results of this methodologies were used to feed the multimedia augmentation layer.

The data warehouse is composed of soccer and basketball data, more precisely data from English Premier League and NBA respectively. ETL and ELT processes were adopted to organize the data from the source systems.

BI and ML were used to analyze the data gathered. Initially BI allowed an exploratory data analysis and identified gaps in the organization of the data warehouse. ML applied the data to forecast the outcomes in both sports, in which the algorithms applied were: decision tree, SVM, random forest, extremely randomized trees, and extreme gradient boosting. Except for the extreme gradient boosting that solved the problem through regression and classification, the other methods solved through classification.

The simulation was applied in basketball to predict game results and enables the simulation play-by-play of a basketball game. The main difference of the simulation is providing visualization of the events of the game, and at the end, it is possible to analyze the impact of the players through the statistics box score.

To enhance the viewer experience, a system was developed to emit a sport live stream, and at the same time display overlays above the video. The overlays presented contain interesting statistical facts about the game transmitted without there being a direct intrusion in the main content.

7.1 Fulfillment of Goals

The goals proposed for this dissertation were achieved with success. Both in the prediction of game outcomes and in the enhancement of the spectator's experience.

In the case of basketball, the NBA season 2015-2016 was predicted applying machine learning techniques and through simulation. In ML, the most successful model was the SVM that reached an accuracy of 67.96%, a value that surpassed some cases of the literature. One feature engineered was the players' weight, which is an attribute not much used in the state of art, normally the focus is on the team values. In the simulation, the model achieved an accuracy of 64.80% and created much value to the state of the art, since there are almost no articles that explain in detail the development of a basketball simulator.

In the case of soccer, the EPL season 2017-2018 was predicted applying machine learning techniques. The most successful model was the extreme gradient boosting regressor with an accuracy of 54.44%, the fact that the value is much lower than in basketball is due to the soccer game has three possible outcomes. A common problem mentioned in the literature is the difficulty in predicting draws, which also happened in this work.

About multimedia augmentation, although there are many products on the market, most handle each sport as a different system. In this dissertation, a prototype was developed that integrates basketball and soccer in the same system, and the generation of overlays presented on the main content is totally dynamic. Whether overlays of teams or players, it is quickly generated for different sports, surpassing the barrier of, for example, the type of statistics being different.

7.2 Future Work

After the conclusion of the dissertation development and consequent analysis of the results, there are several aspects that can be worked out in the future to improve the models and take advantage of their results.

Regarding the game outcome predictions, as seen in the state of the art, there are multiple techniques that could be applied. Markov models are a good example and to implement in a more detailed way, and probably with better results, it would be necessary to include play-by-play data in the data set. About the work developed, in the case of soccer, information about players' weight is not used, therefore a good improvement would be the introduction of player ratings on the models.

Concerning the simulation, beyond the implementation of more game actions (states) in basketball, such as assists and fouls, can also be developed a simulator for soccer. In the case of basketball, a significant improvement would be to increase the data set. Providing statistics of jump ball, tracking data, and play-by-play data to the model would allow the development of new features or update the formulation of certain features. For example, the probability of a player shot considers the defender's position and the pass network be based on the history of passes between teammates. As the simulator's code is structured as a state machine, the implementation of new

states is easy and flexible. Another improvement would be adding parameters as inputs about the game strategy of a team.

About the use of the results, a significant improvement would be gathering the game outcome predictions and insights from models into an interactive dashboard. In this way, the sports organization staff would have more ability and flexibility in reaching the information, without needing a data expert nearby.

References

[1] P. Dizikes, *Big issues on the table at the mit sloan sports analytics conference*, http://news.mit.edu/2019/13th-sloan-sports-analytics-conference-0306 (accessed Jun. 21, 2019).

[2] E. Koteff-Moreano, M. Milward, S. Vasillief, and M. Woodworth, "Technology, media, and telecommunications predictions 2019", Deloitte, Tech. Rep., 2018. [Online]. Available: https://www2.deloitte.com/content/dam/insights/us/articles/TMT-Predictions_2019/DI_TMT-predictions_2019.pdf.

[3] M. Intelligence, *Sports analytics market - growth, trends, and forecasts (2019-2024)*, https://www.mordorintelligence.com/industry-reports/sports-analytics-market (accessed Jun. 05, 2019).

[4] A. De Smet, A. Webb, and J. Luhnow, *How the houston astros are winning through advanced analytics*, https://www.mckinsey.com/business-functions/organization/our-insights/how-the-houston-astros-are-winning-through-advanced-analytics (accessed Feb. 05, 2019).

[5] B. |. A. |. Hamilton, *Big data in the dugout*, https://www.boozallen.com/e/media/company-news/big-data-in-the-dugout.html (accessed Feb. 05, 2019).

[6] S. Cacciola, *Eyes on stats, players hire help to crunch them*, https://www.nytimes.com/2014/05/28/sports/basketball/eyes-on-stats-nba-players-hire-help-to-crunch-them.html (accessed Feb. 05, 2019).

[7] E. Cook, *Percentage baseball*. Waverly Press, 1964.

[8] W. contributors, *Moneyball (film)*, https://en.wikipedia.org/wiki/Moneyball_(film) (accessed Jun. 05, 2019).

[9] B. de Ville, *Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner*. SAS Institute, 2006.

[10] J. Laursen Gert H. N.; Thorlund,
Business analytics for managers: taking business intelligence beyond reporting,
ser. Wiley and SAS business series. John Wiley & Sons, 2017.

[11] R. Kimball and J. Caserta, *The Data Warehouse ETL Toolkit : Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*, 1st ed. 2004.

[12] W. contributors, *Extract, transform, load - wikipedia*,
https://en.wikipedia.org/wiki/Extract,_transform,_load (accessed Jun. 05, 2019).

[13] C. Ballard, V. Gomes, G. Hilz, M. Panthagani, C. Samuelsen, and I. Redbooks,
Data Warehousing with the Informix Dynamic Server, ser. IBM redbooks.
IBM Redbooks, 2009. [Online]. Available:
http://www.redbooks.ibm.com/redbooks/pdfs/sg247788.pdf.

[14] S. B. Journal, *Fan analytics movement reaching more colleges*,
https://www.sportsbusinessdaily.com/Journal/Issues/2015/06/01/Colleges/Paciolan-analytics.aspx (accessed Jun. 29, 2019).

[15] C. K. Bukstein Scott; Harrison, *Sport business analytics: using data to increase revenue and improve operational efficiency*, ser. Data analytics applications.
CRC Press;Auerbach Publications, 2017.

[16] I. B. Data, *New york mets sign sas® to lead off analytics lineup*,
https://insidebigdata.com/2014/10/30/new-york-mets-sign-sas-lead-analytics-lineup/ (accessed Jun. 29, 2019).

[17] K. Goldsberry, "Courtvision : New visual and spatial analytics for the nba",
in *MIT Sloan Sports Analytics Conference*.
[Online]. Available: http://www.sloansportsconference.com/wp-content/uploads/2012/02/Goldsberry_Sloan_Submission.pdf.

[18] O. Caya and A. Bourdon, "A framework of value creation from business intelligence and analytics in competitive sports",
in *2016 49th Hawaii International Conference on System Sciences (HICSS)*,
pp. 1061–1071. DOI: 10.1109/HICSS.2016.136.

[19] T. N. Y. Times, *The no-stats all-star*,
https://www.nytimes.com/2009/02/15/magazine/15Battier-t.html
(accessed Jun. 30, 2019).

[20] S. Few, *Information Dashboard Design: The Effective Visual Communication of Data*,
ser. O'Reilly Series. O'Reilly Media, Incorporated, 2006.

[21] ——, *Information Dashboard Design: Displaying Data for At-a-glance Monitoring*.
Analytics Press, 2013.

[22] I. M. Franks and G. Miller, "Training coaches to observe and remember", *Journal of Sports Sciences*, vol. 9, no. 3, pp. 285–297, 1991. DOI: 10.1080/02640419108729890.

[23] P. Laird and L. Waters, "Eyewitness recollection of sport coaches", *International Journal of Performance Analysis in Sport*, vol. 8, no. 1, pp. 76–84, 2008. DOI: 10.1080/24748668.2008.11868424.

[24] L. H. Peter O'Donoghue, *Data Analysis in Sport*, 1st ed., ser. Routledge Studies in Sports Performance Analysis. Routledge, 2014.

[25] A. G. Losada, R. Therón, and A. Benito, "Bkviz: A basketball visual analysis tool", *IEEE Computer Graphics and Applications*, vol. 36, no. 6, pp. 58–68, 2016. DOI: 10.1109/MCG.2016.124.

[26] H. Pileggi, C. D. Stolper, J. M. Boyle, and J. T. Stasko, "Snapshot: Visualization to propel ice hockey analytics", *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2819–2828, 2012. DOI: 10.1109/TVCG.2012.263.

[27] C. Perin, R. Vuillemot, and J. Fekete, "Soccerstories: A kick-off for visual soccer analysis", *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2506–2515, 2013. DOI: 10.1109/TVCG.2013.192.

[28] Sportradar, *Sports data services by sportradar | digital solutions | integrity | betting*, https://www.sportradar.com/ (accessed Feb. 07, 2019).

[29] ——, *Advanced widgets - information beyond the essentials*, https://www.sportradar.com/media/widgets-and-cards/advanced-widgets/ (accessed Feb. 07, 2019).

[30] ——, *Performance widgets - sports data widgets for nfl football and soccer*, https://www.sportradar.com/media/widgets-and-cards/performance-widgets/ (accessed Feb. 07, 2019).

[31] Dartfish, *Take efficient decisions based on smart video and data solutions.* https://www.dartfish.com/ (accessed Feb. 07, 2019).

[32] ——, *Bring data intelligence to your sport.* https://www.dartfish.com/mobile_s (accessed Feb. 07, 2019).

[33] ——, *Live video performance & data analysis.* https://www.dartfish.com/live_s (accessed Feb. 07, 2019).

[34] FiveThirtyEight, *Nate silver's fivethirtyeight uses statistical analysis — hard numbers — to tell compelling stories about politics, sports, science, economics and culture.* https://fivethirtyeight.com/ (accessed Feb. 07, 2019).

[35] Vizrt, *About vizrt*, https://www.vizrt.com/vizrt (accessed Feb. 07, 2019).

[36] A. Burkov, *The Hundred-Page Machine Learning Book*. Burkov, Andriy, 2019.

[37] S.-l. developers, *Supervised learning*, https://scikit-learn.org/stable/supervised_learning.html#supervised-learning (accessed Jun. 12, 2019).

[38] R. Baboota and H. Kaur, "Predictive analysis and modelling football results using machine learning approach for english premier league", *International Journal of Forecasting*, vol. 35, no. 2, pp. 741–755, 2019. DOI: 10.1016/j.ijforecast.2018.01.003.

[39] P.-F. Pai, L.-H. ChangLiao, and K.-P. Lin, "Analyzing basketball games by a support vector machines with decision tree model", *Neural Computing and Applications*, vol. 28, no. 12, pp. 4159–4167, 2016. DOI: 10.1007/s00521-016-2321-9.

[40] C. Cao, "Sports data mining technology used in basketball outcome prediction", Thesis, 2012. [Online]. Available: https://arrow.dit.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis.

[41] R. Valenzuela, "Predicting national basketball association game outcomes using ensemble learning techniques", PhD thesis, California State University, Long Beach, 2018.

[42] A. Zimmermann, "Basketball predictions in the ncaab and nba: Similarities and differences: Basketball predictions in the ncaab and nba", *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Jun. 2016. DOI: 10.1002/sam.11319.

[43] N. Danisik, P. Lacko, and M. Farkas, "Football match prediction using players attributes", in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, Sep. 2018, pp. 201–206. DOI: 10.1109/DISA.2018.8490613.

[44] B. Ulmer and M. Fernandez, "Predicting soccer match results in the english premier league", 2014.

[45] V. Maini, *Machine learning for humans, part 3: Unsupervised learning*, https://medium.com/machine-learning-for-humans/unsupervised-learning-f45587588294 (accessed Jun. 13, 2019).

[46] X. Wei, P. Lucey, S. Morgan, P. Carr, M. Reid, and S. Sridharan, "Predicting serves in tennis using style priors", *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2207–2215, Jul. 2015. DOI: 10.1145/2783258.2788598.

[47] W. Kenton, *Stochastic modeling*, https://www.investopedia.com/terms/s/stochastic-modeling.asp (accessed Feb. 07, 2019).

[48] R. Serfozo, *Basics of Applied Stochastic Processes, Springer 2009*. 2009. DOI: 10.1007/978-3-540-89332-5.

[49] D. Cervone, A. D'Amour, L. Bornn, and K. Goldsberry, "Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data", in *MIT Sloan Sports Analytics Conference*, 2014. [Online]. Available: http://www.sloansportsconference.com/wp-content/uploads/2018/09/cervone_ssac_2014.pdf.

[50] K. Goldner, *A Markov Model of Football: Using Stochastic Processes to Model a Football Drive*. 2012, vol. 8, pp. 9–9. DOI: 10.1515/1559-0410.1400.

[51] A. Tsokos, S. Narayanan, I. Kosmidis, G. Baio, M. Cucuringu, G. Whitaker, and F. Király, "Modeling outcomes of soccer matches", *Machine Learning*, 2018. DOI: 10.1007/s10994-018-5741-1.

[52] F. J. R. Ruiz and F. Perez-Cruz, "A generative model for predicting outcomes in college basketball", *Journal of Quantitative Analysis in Sports*, vol. 11, no. 1, p. 39, 2015. DOI: 10.1515/jqas-2014-0055.

[53] A. C. Constantinou, "Dolores: A model that predicts football match outcomes from all over the world", *Machine Learning*, 2018. DOI: 10.1007/s10994-018-5703-7.

[54] Kaggle, *Your home for data science*, https://www.kaggle.com/ (accessed Jun. 10, 2019).

[55] K. Team, *Predicting march madness: 1st place winner, zach bradshaw | no free hunch*, http://blog.kaggle.com/2015/04/17/predicting-march-madness-1st-place-finisher-zach-bradshaw/ (accessed Jun. 26, 2019).

[56] ——, *March machine learning mania 2016, winner's interview: 1st place, miguel alomar | no free hunch*, http://blog.kaggle.com/2016/05/10/march-machine-learning-mania-2016-winners-interview-1st-place-miguel-alomar/ (accessed Jun. 26, 2019).

[57] W. contributors, *Computer simulation - wikipedia*, https://en.wikipedia.org/wiki/Computer_simulation (accessed Feb. 07, 2019).

[58] M.-h. Oh, S. Keshri, and G. Iyengar, "Graphical model for baskeball match simulation", in *MIT Sloan Sports Analytics Conference*, 2015. [Online]. Available: http://www.sloansportsconference.com/wp-content/uploads/2015/02/SSAC15-RP-Finalist-Graphical-model-for-basketball-match-simulation.pdf.

[59] P. Vračar, E. Štrumbelj, and I. Kononenko, "Modeling basketball play-by-play data", *Expert Systems with Applications*, vol. 44, pp. 58–66, 2016. DOI: 10.1016/j.eswa.2015.09.004.

[60] Livestrong, *What are the benefits of video analysis in sports?*, https://www.livestrong.com/article/525371-what-are-the-benefits-of-video-analysis-in-sports/ (accessed Feb. 05, 2019).

[61] R. Wood, *Video analysis in sports*, https://www.topendsports.com/biomechanics/video-analysis.htm (accessed Feb. 05, 2019), 2010.

[62] H. Shih, "A survey of content-aware video analysis for sports", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1212–1231, 2018. DOI: 10.1109/TCSVT.2017.2655624.

[63] H.-T. Chen, C.-L. Chou, T.-S. Fu, S.-Y. Lee, and B.-S. P. Lin, "Recognizing tactic patterns in broadcast basketball video using player trajectory", *Journal of Visual Communication and Image Representation*, vol. 23, no. 6, pp. 932–947, 2012. DOI: 10.1016/j.jvcir.2012.06.003.

[64] L.-H. Chen, C.-W. Su, and H.-A. Hsiao, *Player trajectory reconstruction for tactical analysis*. 2018, vol. 77. DOI: 10.1007/s11042-018-6164-5.

[65] W. Hobbs, A. D. Gorman, S. Morgan, M. Mooney, and J. Freeston, *Measuring spatial scoring effectiveness in women's basketball at the 2016 Olympic Games*. 2018, vol. 18, pp. 1–13. DOI: 10.1080/24748668.2018.1550892.

[66] K. Przednowek, T. Krzeszowski, K. H. Przednowek, and P. Lenik, *A System for Analysing the Basketball Free Throw Trajectory Based on Particle Swarm Optimization*. 2018, vol. 8, p. 2090. DOI: 10.3390/app8112090.

[67] A. Atrish, N. Singh, and V. Kumar, "Enhanced homography-based sports image components analysis system: Aicc 2018", in. 2019, pp. 495–505. DOI: 10.1007/978-981-13-1580-0_48.

[68] I. Atmosukarto, B. Ghanem, M. Saadalla, and N. Ahuja, *Recognizing Team Formation in American Football*. 2014, vol. 71, pp. 271–291. DOI: 10.1007/978-3-319-09396-3_13.

[69] C. Chen and L. Chen, "Novel framework for sports video analysis: A basketball case study", in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 961–965. DOI: 10.1109/ICIP.2014.7025193.

[70] X. Yu and W. Ding, *Game suspension boundary detection by reading two clocks for broadcast basketball video*. 2015, pp. 1–4. DOI: 10.1145/2808492.2808537.

[71] H. Liu, S. Jiang, Q. Huang, and C. Xu, *A generic virtual content insertion system based on visual attention analysis*. 2008, pp. 379–388. DOI: 10.1145/1459359.1459410.

[72] X. Yu, X. Yan, T. T. Phuong Chi, and L. F. Cheong,
Inserting 3d projected virtual content into broadcast tennis video, Conference Paper, 2006. DOI: 10.1145/1180639.1180767.

[73] S. Monji-Azad, S. Kasaei, and A. Eftekhari-Moghadam, "An efficient augmented reality method for sports scene visualization from single moving camera",
in *2014 22nd Iranian Conference on Electrical Engineering (ICEE)*, pp. 1064–1069.
DOI: 10.1109/IranianCEE.2014.6999693.

[74] Y. Li, K. W. Wan, X. Yan, and C. Xu,
Real time advertisement insertion in baseball video based on advertisement effect, Conference Paper, 2005. DOI: 10.1145/1101149.1101221.

[75] C. Xu, K. W. Wan, S. H. Bui, and Q. Tian,
"Implanting virtual advertisement into broadcast soccer video",
in *Advances in Multimedia Information Processing - PCM 2004*,
K. Aizawa, Y. Nakamura, and S. Satoh, Eds., Springer Berlin Heidelberg, pp. 264–271.

[76] Dartfish, *A 3d graphic rendering solution for sports video analysis.*
https://www.dartfish.com/pro_s (accessed Feb. 07, 2019).

[77] ——, *Optimise your athlete's performance.*
https://www.dartfish.com/360 (accessed Jun. 05, 2019).

[78] Vizrt, *Sports analysis solution*,
https://www.vizrt.com/sports/sports-analysis (accessed Jun. 05, 2019).

[79] S. R. LLC, *Basketball-reference.com - basketball statistics and history*,
https://www.basketball-reference.com (accessed Jun. 05, 2019).

[80] Python, *Welcome to python.org*,
https://www.python.org/ (accessed Jun. 08, 2019).

[81] J. Bradley, *Nba stats api via basketball reference*,
https://github.com/jaebradley/basketball_reference_web_scraper
(accessed Jun. 06, 2019).

[82] *Beautiful soup documentation — beautiful soup 4.4.0 documentation*,
https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed Jun. 06, 2019).

[83] T. Basham, *Ratelimit*,
https://github.com/tomasbasham/ratelimit (accessed Jun. 06, 2019).

[84] PostgreSQL, *The world's most advanced open source database*,
https://www.postgresql.org/ (accessed Jun. 08, 2019).

[85] *Python data analysis library — pandas: Python data analysis library*,
https://pandas.pydata.org/ (accessed Jun. 08, 2019).

[86] Psycopg, *Postgresql database adapter for python — psycopg 2.8.3.dev0 documentation*, http://initd.org/psycopg/docs/ (accessed Jun. 08, 2019).

[87] OptaSports, *World leaders in sports data*, https://www.optasports.com/ (accessed Jun. 10, 2019).

[88] ShubhamPawar, *English premier league in-game match data | kaggle*, https://www.kaggle.com/shubhmamp/english-premier-league-match-data (accessed Jun. 08, 2019).

[89] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, and G. Louppe, *Scikit-learn: Machine Learning in Python*. 2012, vol. 12.

[90] R. Pierre, *Data leakage, part i: Think you have a great machine learning model? think again*, https://towardsdatascience.com/data-leakage-part-i-think-you-have-a-great-machine-learning-model-think-again-ad44921fbf34 (accessed Jun. 26, 2019).

[91] Basketball-Reference, *Glossary*, https://www.basketball-reference.com/about/glossary.html (accessed Jun. 17, 2019).

[92] S.-l. developers, *Support vector machines — tips on practical use*, https://scikit-learn.org/stable/modules/svm.html#tips-on-practical-use (accessed Jun. 26, 2019).

[93] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[94] A. Constantinou and N. Fenton, *Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries*. 2013, vol. 9, pp. 37–50. DOI: 10.1515/jqas-2012-0036.

[95] ESPN, *Nba: Projected standings for 2015-16*, https://www.espn.com/nba/story/_/id/13980821/projected-standings-2015-16 (accessed Jun. 26, 2019).

[96] P. BI, *Interactive data visualization bi tools*, https://powerbi.microsoft.com/en-us/ (accessed Jun. 23, 2019).

[97] Tableau, *Business intelligence and analytics software*, https://www.tableau.com/ (accessed Jun. 23, 2019).

[98] Qlik, *Data analytics for modern business intelligence*, https://www.qlik.com/us (accessed Jun. 23, 2019).

[99] B. Survey, *Business intelligence, analytics & planning tool reviews*, https://bi-survey.com (accessed Jun. 23, 2019).

[100] M. Rouse, *What is restful api? - definition from whatis.com*, https://searchmicroservices.techtarget.com/definition/RESTful-API (accessed Jun. 26, 2019).

[101] A. Ronacher, *Flask - a python microframework*, https://http://flask.pocoo.org/ (accessed Jun. 26, 2019).

[102] *Video.js: The player framework*, https://videojs.com/ (accessed Jun. 26, 2019).